STAT 961: Statistical Methodology

Fall 2021

Course and Instructor Information

Course information

Classroom: TBA
Class time: Tue/Thu 10:15-11:45am
Website: https://canvas.upenn.edu/courses/1597407

Instructor: Eugene Katsevich
Office: 311 Academic Research Building
Email: ekatsevi@wharton.upenn.edu
Office hours: Tue 1:00-3:00pm

Teaching assistant: Hua Wang
Office: TBA
Email: wanghua@wharton.upenn.edu
Office hours: TBA

Course Description

This 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 961is a fast-paced, PhD level course that requires significant preparation. Students are expected to have the following prerequisites:

- *Linear algebra* at the level of MATH 312 (including bases, vector spaces, inner products, orthogonal projections, and matrix decompositions)
- *Probability* at the level of STAT 430 (including random variables, probability distributions, multivariate normal random variables, and the central limit theorem)
- *Statistical inference* at the level of STAT 431 (including hypothesis testing, p-values, confidence intervals, and maximum likelihood estimation)
- *Programming* experience in R

List of topics

(tentative and subject to change)

Linear models

- Lease squares estimation: normal equations, geometric interpretation via projections, Gauss-Markov theorem, linear model examples, decomposition of variance, orthogonalization, partial correlation
- Inference in linear models: Normal random variables, hypothesis testing, confidence intervals, prediction intervals, power, collinearity
- Model-checking, model misspecification, and robust alternatives: Residual plots, leverage and influence, Huber-White estimator, pairs bootstrap, permutation tests, Wilcoxon test

Generalized linear models

- Exponential family distributions
- Maximum likelihood estimation for GLMs and iteratively reweighed least squares
- Testing and estimation in GLMs, deviances, goodness of fit
- Special cases of GLMs: Logistic regression, multinomial logistic regression, Poisson regression, negative binomial regression

Additional topics

- Linear mixed models
- Multiple testing
- Penalized regression and cross-validation

Assignments and Exams [needs update]

Assessment is based on homework, a midterm exam, and a final exam.

Homework (35%)

There will be four homework assignments, one at the end of each of the first four units. These homework assignments will involve conceptual questions as well as R programming questions. Students can work in teams of up to three people. Submission will be through Gradescope; one per team. Students can use Piazza to find team members. **Each student will get three "free" late days for homework submission over the course of the semester, for which there will be no late penalty. Late passes must be claimed through Canvas.** 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 measured in whole days.

Midterm exam (30%)

The midterm exam will take place on Monday, March 22 from 6-8pm. This individual work, open book exam will involve conceptual questions as well as R programming questions.

Final exam (25%)

In the final project, students will apply the methods they learned in class to tackle data mining problems of personal interest to them. Working in teams of up to three, students will identify an analysis goal and either collect or find a relevant dataset. **The final project report (maximum**

15 pages) is due on Sunday, May 2 by 11:59pm. Selected teams will be invited to showcase their projects at the Data Science Live event on Friday, April 30 (see e.g. DSL from Fall 2019).

Course Grades

Course grades will be assigned based on the following ranges:

Course Materials

Textbook

Our required textbook is known as ISLR and is freely available online:

- Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*, First Edition, 2013, Springer New York.

Laptops

Laptops are required for hands-on data analysis, an essential part of the course.

Software

Students must download the free statistical computing language R and the integrated development environment RStudio.

Also LaTeX

Course Policies

Make-ups and extensions [needs update]

Instead of offering make-ups or deadline extensions on a case-by-case basis, most assignments grading policies (see "Assignments and Exams" section above) have built-in buffer mechanisms to accommodate unforeseen circumstances. No additional make-ups or deadline extensions will be granted, unless extenuating circumstances arise.

Regrades [needs update]

Grading of all non-multiple-choice assignments, i.e. homework, midterm, and final project, will be through Gradescope. In this system, 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. Grades will not necessarily stay the same or increase as a result of a regrade request.

Academic integrity [needs update]

Students are expected to be familiar with and comply with Penn's Code of Academic Integrity, which is available online. I generally have a zero-tolerance policy for cheating, and all violations will result in substantial penalties. If you have any doubts or questions about what constitutes academic misconduct, please do not hesitate to contact me.

COVID-related accommodations

Office hour scheduling. During the first week of class, the instructor will survey students' availability to attend office hours. Office hours will then be scheduled to make sure there is at least one office hour time during the week that works for each student.

In-class assessments. Entry tickets, exit tickets and quizzes are normally submitted during class time. Each student will fall into one of three categories: (1) can attend the section they signed up for, (2) can't attend the section they signed up for but can attend the other section, or (3) can't attend either section. During the first week of class, the instructor will survey students' availability to attend class sections. Students in category 2 will be allowed to attend the other section, and an alternative time for students in category 3 to submit their "in-class" assessments will be arranged. **Students who do not respond to the survey will be assumed to be able to attend the section they signed up for and will be expected to submit their in-class assessments during that section.**

Midterm exam. During the first week of class, the instructor will survey students' availability to take the midterm exam at the appointed time (on March 22 from 6-8pm EST). Those students whose time zone precludes them from taking the midterm at the appointed time will be given an alternative time slot. **Students who do not respond to the survey will be assumed to be able to take the midterm at the appointed time.**

Technical difficulties. Late penalties on assignments will be waived for students who encounter technical difficulties with submission, provided they take a screenshot of the issue and email the instructor as soon as possible.

Accessibility

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 by February 1. The instructor will then work with the student and SDS to provide reasonable accommodations.