36-402/608 Advanced Methods for Data Analysis

Course Policies and Syllabus, Spring 2021

Instructor: Ann Lee (annlee@nadrew.cmu.edu) and Alex Reinhart (areinhar@stat.cmu.edu)
Office hours: TBA

Teaching assistants: TBA
Office hours: TBA (The head TA will also run a Q&A discussion board on Piazza but this board is *not* a replacement of office hours, and only meant for short clarifications and questions that can be answered with a few sentences.)

Lectures: Tuesdays and Thursdays 10:40am–noon Eastern time, via Zoom. Recordings will be available via Canvas after class, though live attendance is strongly encouraged.

Textbook:
Shalizi, *Advanced Data Analysis from an Elementary Point of View*, September 2019 (posted on Canvas; page references will refer to this version)

Recommended additional texts:
1. Long & Teetor, *R Cookbook*, 2nd ed., O'Reilly, 2019. (Free e-copy available at https://rc2e.com/)
2. James, Witten, Hastie, and Tibshirani, *An Introduction to Statistical Learning*, Springer 2015. (Free e-copy available at https://statlearning.com/)
3. Wasserman, *All of Statistics*, Springer 2004. (Free e-copy available at https://link.springer.com/book/10.1007/978-0-387-21736-9)

Course website: https://canvas.cmu.edu/
Lecture outlines, lecture check-points, homeworks, and graded assignments will be posted on Canvas.

Lecture Notes: Download outlines from Canvas and complete them during class time.

Lecture Check-Points: Complete the check-points after each lecture, and submit your answers via Canvas *before* the beginning of the following lecture.

Note that Sections A and B meet together and have identical assignments and grading policies; there is no difference between sections.

Course Overview and Learning Objectives

You have learned to use linear regression as your main tool in data analysis. So what is next? This course aims to train you to use advanced statistical methods for analyzing data.

We start with the linear model and build on the theory and applications that you have already seen in this setting in order to explore richer model classes, more kinds of data, and more complex setups. All the while, our intention is to develop an intuitive understanding of the methods and their limitations, a formal understanding of the same concepts, and the practical/programmatic skills to apply these methods in real problems. Upon completing this course, you should be able to tackle new applied statistics problems, by:

- selecting the appropriate techniques and justifying your choices
- learn the basic mathematical theory underlying these models
- implementing these techniques programmatically (using R) and critically evaluating your results
- explaining your results to your collaborators and researchers outside of statistics.

Prerequisites

The formal prerequisite is the *linear regression* part of 36-401: Modern Regression. In addition, we will assume that you are familiar with basic probability, statistics, linear algebra, and R programming. Specifically, here is a list of topics that you are expected to know already.[1]

- *Probability.* Event, random variable, indicator variable; probability mass function, probability density function, cumulative distribution function; joint and marginal distributions; conditional probability, Bayes's rule; independence; expectation, variance; binomial, Poisson, Gaussian distributions.
- *Statistics.* Sampling from a population; mean, variance, standard deviation, median, covariance, correlation, and their sample versions; histogram; likelihood, maximum likelihood estimation; point estimates, standard errors, confidence intervals, p-values; linear regression, response and predictor variables, coefficients, residuals.
- *Linear algebra.* Vectors and scalars; components of a vector, geometry of vectors; vector arithmetic: adding vectors, multiplying vectors by scalars, dot product of vectors; coordinate basis, change of basis; matrices, matrix arithmetic: matrix addition, matrix multiplication, matrix inversion, multiplication of matrices and vectors; eigenvalues and eigenvectors of a matrix.
- *R programming.* R arithmetic (scalar, vector, and matrix operations); writing functions; reading in data sets, using and manipulating data structures; installing, loading, and using packages; plotting.

---

[1] For a more complete list, see "Concepts You Should Know" in Shalizi (Sept 2019) page 14.

Homework

There will be 11–12 weekly homework assignments **due on Fridays by 3:00 pm**, except for around exam times and when otherwise stated. Homeworks should be turned in electronically using Canvas. You are encouraged to back up your work by uploading your work as you complete problems (this will guard you against receiving a zero on an assignment because of some unforeseeable circumstances); only your latest submitted version will be graded.

NO LATE HOMEWORKS WILL BE ACCEPTED FOR ANY REASON.

Assignments, solutions and your graded assignments will be posted on Canvas. The assignments will be posted a week before the due date.

You may have to read ahead to do some of the problems. You are allowed to discuss the assignments with other students in the course, but the work that you hand in, both written work and code, **must be your own** and written up independently.

You should submit your homework through Canvas *in two parts* with

1. *Writeup.* An important part of the learning the data analysis trade is learning how to communicate. Prepare a writeup to the homework questions and work hard to make your submission as readable as you possibly can; making sure figures are of proper (large) size with labeled, readable axes; and so forth.

2. *Code.* Always include your R code for the homework. You should demarcate sections of the code that correspond to the different homework problems using clearly visible comments (e.g., as in `##### Problem 1 #####`).

If you use a knitting program (such as R Markdown `http://rmarkdown.rstudio.com` or knitr `https://yihui.name/knitr/`), then this means that you should submit *both* the knitted file *and* the source that was used to create the knitted file. There will be a 20% point deduction if you only submit a write-up but no code, or if you submit the source of a knitting program but not the knitted file itself.
Note that if only the correct answer is provided, but no relevant derivations, then zero points will be awarded. Some advice: With each exercise, look at your solution, and ask yourself: "Suppose I had been provided in advance with the correct answer to this exercise. Would it be clear to the grader that I understand how to reach that answer?"

Lecture Check-Points

After each lecture and due before the next lecture, you will be asked to answer several questions about the material covered in lecture. These will be posted on Canvas and will

contribute to your course grade. The purpose of these questions is to help you self-assess your understanding of the lecture material. If you answer a question incorrectly, you should immediately try to understand why you got it wrong. I cannot emphasize enough how valuable these lecture checkpoints can be in providing you and me with immediate feedback on how well you understand the course material. I highly recommend engaging in these exercises.

Exams and Grading Policy

All exams are cumulative and you are not allowed to discuss the content of the exams with other students until the solutions have been posted. The dates of the exams will not be moved so please schedule job interviews and other extra-curricular activities around them.

In-Class Tests: There will be two in-class tests during the semester, on **Tuesday, February 18** and **Thursday, March 26**.

There will not be any makeup exams.

Final Exam: There will not be a final exam.

Data Analysis Exams: There will be two take-home data analysis exams during the semester; each will be posted on a Friday in lieu of a regular homework assignment, and due the following **Friday by 3 pm**. Your analysis and results will be turned in as a structured report. We will cover the specifics in detail later in the course.

Final Grades: Final scores will be calculated based on a weighted averages of your scores for the homeworks (20%), the two tests (15% each), the two data analysis exams (20% each), and the lecture checkpoints (10%). Your two lowest homework scores will be dropped. Each homework will receive equal weight, regardless of the number of points.
Grades will be computed using the usual scale: 90% and up guarantees an A, 80% to 90% guarantees a B, and so on. I *may* adjust this scale in your favor to, for example, account for border-line cases or curve the class if needed.

Lectures

You are expected to attend class and actively take notes. I will prepare lecture "outlines" for you to complete during class; these outlines will be posted by 6 pm on the day before the lecture. You are responsible for printing out the handouts and bringing them to class.

Academic Integrity and Plagariasm

All students are expected to comply with the CMU policy on academic integrity:
`https://www.cmu.edu/policies/student-and-student-life/academic-integrity.html`
Cheating, copying, etc. will not be tolerated. Please ask if you are unsure of whether or not your actions are complying with assignment/exam instructions.
Note for both sections of 36-402: *"Any use of solutions provided for any assignment in this course in previous years is strictly prohibited, both for homework and for exams. This prohibition applies even to students who are re-taking the course. Do not copy the old solutions (in whole or in part), do not 'consult' them, do not read them, do not ask your friend who took the course last year if they 'happen to remember' or 'can give you a hint'. Doing any of these things, or anything like these things, is cheating, it is easily detected cheating, and those who thought they could get away with it in the past have failed the course. Even more importantly: doing any of those things means that the assignment doesn't give you a chance to practice; it makes any feedback you get meaningless; and of course it makes any evaluation based on that assignment unfair."*

Accommodations for Students with Disabilities

If you have a disability and are registered with the Office of Disability Resources, I encourage you to use their online system to notify me of your accommodations and discuss your needs with me as early in the semester as possible. I will work with you to ensure that accommodations are provided as appropriate. If you suspect that you may have a disability and would benefit from accommodations but are not yet registered with the Office of Disability Resources, I encourage you to contact them at access@andrew.cmu.edu.

Other Policies

Please note each of the following.

1. It is assumed that you check your Andrew email at least once per day. We will use email and Canvas announcements to communicate important details about the course, so ensure your Canvas settings send you automatic notifications when new announcements are made.

2. Students are expected to be attentive during lecture.
   LAPTOP USE SHOULD PERTAIN DIRECTLY TO THE CLASS.
   All cellphones (or anything else that makes noise) should be silenced during class.

3. Sending email to your professor or teaching assistants should be treated as professional communication. Questions posed by email must be sent at least 24 hours before the time the assignment is due in order to receive a response.

4. No student may record or tape any classroom activity without the consent of the instructor. If a student believes that he/she is disabled and needs to record or tape classroom activity, he/she should contact the Office of Equal Opportunity Services, Disability Resources to request appropriate accommodation.

Study Tips

1. **Attend class and actively take notes.** The professor may not type up or write down everything that is said in class. Reading someone else's notes may also not give you a good idea of what was emphasized in class and the order in which things were written.

2. After each lecture, go over your notes.

   - Highlight or make marginal notes for important words or concepts. Fill in gaps with extra explanations. Study the corresponding sections in the book.
   - Re-do examples yourself, step by step, with pencil and paper. Examples often look easy when explained in class, but often turn out to be much harder when you try them yourself.
   - Write down questions about things you do not understand. Bring these questions to me or the TA and ask them.

3. **DO HOMEWORK PROBLEMS** (even if your lowest scores are dropped). Actively doing problems is the *only* way to learn the material. Try to do the problems yourself before discussing them with other people.

4. Review solutions to assignments even if you received a full score.

5. *Take advantage of office hours and use them productively.* The more specific your question and the more documentation of your attempted solution to homework assignments, the better we will be able to help you.

6. Come to each class with a good knowledge of the material that was covered in the previous class.

Finally, take care of yourself. Do your best to maintain a healthy lifestyle this semester by eating well, exercising, avoiding drugs and alcohol, getting enough sleep and taking some time to relax. This will help you achieve your goals and cope with stress.
If you or anyone you know experiences any academic stress, difficult life events, or feelings like anxiety or depression, we strongly encourage you to seek support. Counseling and Psychological Services (CaPS) is here to help: call 412-268-2922 and visit `https://www.cmu.edu/counseling/`. Consider reaching out to a friend, faculty or family member you trust for help getting connected to the support that can help.

Tentative Course Schedule[2]

[2]**The dates of the take-home and in-class exams are fixed and will not be changed.** The rest of the schedule may vary depending on time and class interests.

Beyond Simple Linear Regression

(1) Week of 2/1	Introduction and regression
	The truth about linear regression
(2) Week of 2/8	Breaking assumptions in linear regression
	Association versus causation; counterfactuals
(3) Week of 2/15	Association, causation, and counterfactuals (cont'd)
	Simpson's paradox
(4) Week of 2/22	**Break day; no classes (Tuesday, February 23)**
	Prediction error
	Cross-validation
(5) Week of 3/1	The bootstrap
	Bootstrap (cont'd)
(6) Week of 3/8	**In-Class Test One (Tuesday, February 18)**
	Bootstrapping regression models
(7) Week of 3/15	Regression and smoothing splines
	Splines (cont'd)
(8) Week of 3/22	Effective degrees of freedom
	Effective degrees of freedom (cont'd)
(9) Week of 3/29	TODO
(10) Week of 4/5	Generalized cross-validation; kernel regression revisited
	Kernel regression (cont'd)
(11) Week of 4/12	Review session
	In-Class Test Two (Thursday, March 26).
	Data Exam One out (Friday, March 27)
(12) Week of 4/19	Additive models
	Direct inference with linear smoothers
	Data Exam One due (Friday, April 3)
(13) Week of 4/26	Logistic regression
	Generalized linear models
(14) Week of 5/3	Generalized linear models (cont'd)
	Spring carnival; no class on Thursday.
(15) Week of TODO	Model checking and goodness of fit
	Model checking (cont'd); review
	Data Exam Two out (Friday, April 24)

Unsupervised Learning and Special Topics

(16) Week of 4/27	Nonparametric density estimation. Special topics.
	Data Exam Two due (Friday, May 1)

Finals week	*No final exams or assignments*