36-402/608 Advanced Methods for Data Analysis

Course Policies and Syllabus, Spring 2022

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

Teaching assistants: TBA
Office hours: TBA (The head TAs 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 8:35–9:55am Eastern time, in GHC 4401 (Section A) and DH 2210 (Section B). Please attend the section you are enrolled for. In-person attendance is expected. Class will meet via Zoom when required by University policy; Zoom links and recordings will in that case be available via Canvas, though live attendance is expected even when the course is remote.

Textbook: Shalizi, *Advanced Data Analysis from an Elementary Point of View*, March 2021 (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 are scheduled in different rooms due to capacity limits, but have identical assignments and grading policies. Professor Reinhart will lecture for Section A and Professor Lee will lecture for Section B. You should attend the section you are enrolled for. When the course is taught online due to University policy, both sections will meet in the same Zoom room.

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.

[1] For a more complete list, see "Concepts You Should Know" in Shalizi (Mar 2021) page 15.

- *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. We recommend taking 36-350 Statistical Computing, or having similar programming experience, before taking this course.

Homework

There will be 11–12 weekly homework assignments **due on Fridays by 3:00 pm** Eastern Time, except for around data exam times and when otherwise stated. Homeworks must be turned in electronically using Gradescope. 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. But note that your lowest scores will be dropped—see details in the grading policy below.

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 Gradescope as a single PDF file that includes

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, preferably *embedded* in your write-up as R Markdown "code chunks".

We strongly recommend that you use a knitting program (such as R Markdown `http://rmarkdown.rstudio.com` using knitr `https://yihui.name/knitr/`) which will allow you to embed your code into your write-up as "code chunks". If you instead include your code at the end of your write-up, then you will need to demarcate sections of the code that correspond to the different homework problems using clearly visible comments (e.g., as in `##### Problem 1 #####`).

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 PDF 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. We cannot emphasize enough how valuable these lecture checkpoints can be in providing you and us with immediate feedback on how well you understand the course material. We 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 graded in-class tests; dates are given in the course schedule below. Note that we reserve the right to cancel the in-class tests and adjust the grade percentages accordingly, if COVID makes in-class exams difficult.

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 Eastern time**. (See last page of syllabus for dates; if you need special accommodations because of disabilities or religious holidays then please notify us as soon as possible.) 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 (30%), the two tests (10% each), the two data analysis exams (20% each), and the lecture checkpoints (10%). **Your two lowest homework scores** will be dropped.

Grades will be computed using the usual scale: 90% and up guarantees an A, 80% to 90% guarantees a B, and so on. We *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; in-person lectures will not be recorded. We will prepare lecture "outlines" for you to complete during class; these outlines will be posted on Canvas by 6 pm Eastern Time on the day before the lecture.

When courses are taught online due to University policy, we will record lectures and make them available to students via Canvas. These recordings are **not to be redistributed** to anyone outside the course without the written permission of the instructors.

Academic Integrity and Plagiarism

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.

Important warning: *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, we encourage you to use their online system to notify us of your accommodations and discuss your needs with us as early in the semester as possible. We 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, we 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; **please ensure your Canvas settings send you automatic notifications when new announcements are made.**

2. Students are expected to attend and be attentive during lecture.
   When on Zoom, please mute yourself during lecture. You are encouraged to ask questions during class by using the Zoom raise hand feature (preferred) or by typing in questions into the Zoom chat (which we will try to monitor at regular intervals).

3. Posting questions on Piazza should be treated as professional communication. Questions must be posted at least 24 hours before the time the assignment is due in order to receive a response.

Study Tips

1. **Attend class and actively take notes as well as participate in in-class polls**. 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.

   - Fill in gaps with extra explanations. Study the corresponding sections in the book.
   - Re-do examples yourself, step by step, with pencil and paper. Explain each step and decision to yourself. 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 us or the TAs 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.

Diversity and Inclusion

We must treat every individual with respect. We are diverse in many ways, and this diversity is fundamental to building and maintaining an equitable and inclusive campus community. Diversity can refer to multiple ways that we identify ourselves, including but not limited to race, color, national origin, language, sex, disability, age, sexual orientation, gender identity, religion, creed, ancestry, belief, veteran status, or genetic information. Each of these diverse identities, along with many others not mentioned here, shape the perspectives our students, faculty, and staff bring to our campus. We, at CMU, will work to promote diversity, equity and inclusion not only because diversity fuels excellence and innovation, but because we want to pursue justice. We acknowledge our imperfections while we also fully commit to the work, inside and outside of our classrooms, of building and sustaining a campus community that increasingly embraces these core values.

Each of us is responsible for creating a safer, more inclusive environment.

Unfortunately, incidents of bias or discrimination do occur, whether intentional or unintentional. They contribute to creating an unwelcoming environment for individuals and groups at the university. Therefore, the university encourages anyone who experiences or observes unfair or hostile treatment on the basis of identity to speak out for justice and support, within the moment of the incident or after the incident has passed. Anyone can share these experiences using the following resources:

- Center for Student Diversity and Inclusion: csdi@andrew.cmu.edu, (412) 268-2150
- Report-It online anonymous reporting platform: reportit.net username: tartans password: plaid

All reports will be documented and deliberated to determine if there should be any following actions. Regardless of incident type, the university will use all shared experiences to transform our campus climate to be more equitable and just.

Tentative Course Schedule[2]

(1) Week of 1/17	Introduction and regression
(2) Week of 1/24	The truth about linear regression
	Breaking assumptions in linear regression
(3) Week of 1/31	Association, causation, and counterfactuals
	Simpson's paradox. Prediction error.
(4) Week of 2/7	Prediction error (cont'd)
	Cross-validation
(5) Week of 2/14	The Bootstrap
	Bootstrap (cont'd)
(6) Week of 2/21	Bootstrapping regression models
	Regression and smoothing splines
(7) Week of 2/28	Splines (cont'd)
	Effective degrees of freedom
(8) Week of 3/7	**Spring break: no classes**
(9) Week of 3/14	Effective degrees of freedom (cont'd)
	In-Class Test One (Thursday, March 17)
	Data Exam One out (Friday, March 18)
(10) Week of 3/21	Generalized cross-validation; kernel regression revisited
	Kernel regression (cont'd)
	Data Exam One due (Friday, March 25)
(11) Week of 3/28	Additive models
	Direct inference with linear smoothers
(12) Week of 4/3	Logistic regression
	Spring Carnival: no classes (Thursday, April 7)
(13) Week of 4/11	More on logistic regression
	Generalized linear models
(14) Week of 4/18	Generalized linear models (cont'd)
	Model checking and goodness of fit
(15) Week of 4/25	Model checking (cont'd)
	Nonparametric density estimation
	In-Class Test Two (Thursday, April 28)
	Data Exam Two out (Friday, April 29)
Finals week	**Data Exam Two due (Friday, May 6**)

[2]**The dates of the take-home and in-class exams are fixed.** The rest of the schedule may vary depending on time and class interests.