36-401/607 Modern Regression

Course Policies and Syllabus, Fall 2023

Instructor: Alex Reinhart (areinhar@stat.cmu.edu)
Office hours: To be posted on Canvas

Teaching assistants: Head TA Akshay Prasadan, and a team of graduate TAs and undergraduate graders
Office hours: Schedule to be posted on Canvas. (The TAs will also run a Q&A discussion board on Piazza but this 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 11:00am–12:20pm Eastern time in the Mellon Institute Auditorium. **See Canvas for directions**

Textbook: Sanford Weisberg, Applied Linear Regression, 4th Edition, 2013. (Available online through the CMU Library.)

Additional references:
1. Long & Teetor, *R Cookbook*, 2nd ed., O'Reilly, 2019. `https://rc2e.com/`
2. Wasserman, *All of Statistics*, Springer 2004. (Free e-copy available at `https://link.springer.com/book/10.1007/978-0-387-21736-9`)
3. Wickham, Çetinkaya-Rundel, and Grolemund. *R for Data Science*, 2nd ed., O'Reilly. `https://r4ds.hadley.nz/`

Course website: `https://canvas.cmu.edu/`
Lecture outlines, homeworks, and course announcements will be posted on Canvas. Ensure Canvas is configured to notify you so you receive announcements about the class.

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

Course Overview

This course is an introduction to applied data analysis using linear regression modeling. We will explore real data sets, examine various models for the data, assess the validity of their assumptions, and determine which conclusions we can make (if any). We will use the R programming language to implement our analyses and produce graphs and tables of results. Data analysis is a bit of an art; there may be several valid approaches. We will strongly emphasize the importance of critical thinking about the data and the question of interest. Our overall goal is to use data and a basic set of modeling tools to answer substantive questions, and to present the results in a scientific report.

The course includes a review and discussion of exploratory methods, informal techniques for summarizing and viewing data. We then consider simple linear regression, a model that uses

only one predictor. After briefly reviewing some linear algebra, we turn to multiple linear regression, a model that uses multiple variables to predict the response of interest. For all models, we will examine the underlying assumptions. More specifically, do the data support the assumptions? Do they contradict them? What are the consequences for inference? Finally, we will explore extra topics such as, for example, generalized linear models, additive models, and high-dimensional regression.

A minimum grade of C in any one of the pre-requisites (36-326, 36-226, or 36-236) or a B in 36-218 (for those not pursuing a Statistics major or minor) is required, along with a passing grade in 21-240, 21-241, or 21-242. A grade of C in 36-401 is required to move on to 36-402 or any 36-46x course.

This course uses the free statistical programming language R. We recommend that you take 36-350 Statistical Computing before or at the same time as this course.

Learning Objectives

By the end of the course, students should be able to:

- Use exploratory data analysis tools (e.g., graphical displays) to explore datasets and select appropriate models to answer research questions.
- Fit linear regression models to data, evaluate their assumptions, and interpret their results.
- Apply the basic mathematical theory underlying linear regression models.
- Discuss conclusions and limitations of statistical evidence, in the form of a scientific report.
- Effectively use R, a widely-used statistical package, in data analysis.

Homework

There will be about 10 weekly homework assignments **due on Fridays by 1:00 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. Instead, your lowest homework score will be dropped, so when an illness or family emergency arises, you can decide not to finish that week's assignment without penalty.

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

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. When you collaborate with another student, please write: "I worked with ..." on this assignment. You are not allowed to copy or "consult" solutions from previous years or another section. You are also not allowed to share course solutions with others, even after the end of the course.

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 labels, legends, 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".

There will be a 10-point deduction if you only submit a write-up but no code, or if your submission is poorly formatted and difficult for the graders to read.

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?"

Your submissions will be graded on Gradescope. If you feel a problem has been incorrectly graded, you may submit a regrade request (using the button on Gradescope) up to one week after grades for that assignment have been released. Regrade requests will not be accepted more than one week after grades are released.

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 three 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.

There will not be any makeup exams. We will only count your two highest in-class test scores.

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 1 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.

Final Grades: Final scores will be calculated based on a weighted average of your homework (20%) with the lowest homework score dropped, two out of three in-class tests (20% each), and the two data analysis take-home exams (20% each).

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`

You may use external resources to help with your work, but they must *assist* your own thinking, not replace it. For example, you can use Stack Overflow to help figure out why your R code doesn't work, **but** you must clearly indicate in your submission which parts have come from the outside source, and your overall solution must be your own. Similarly, you can use tools like ChatGPT to explain things to you or suggest improvements to your writing, **but** you cannot simply ask ChatGPT to do the homework for you. If you're unsure if you're allowed to use a resource in a specific way, please ask us first.

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.

As a corollary, posting solutions from this course online (on sites like Course Hero and Chegg) or otherwise sharing them with future students is also an academic integrity violation, and is considered "unauthorized assistance" under CMU's policies.

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. 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`.

Getting Help

There are several main ways to get help with the content of this course:

Office hours These are the best way to get in-depth help with course content and assignments. You are welcome to attend even if you do not have specific questions—you can work on your homework and ask questions as they arise.

Piazza This is available through Canvas and allows you to post questions and answers visible to other students. Questions on Piazza should be treated as professional communication. Post your homework questions at least 24 hours before deadline in order to receive a response.

Email Questions about course content should be asked in office hours or on Piazza. Email should be reserved for individual course logistics questions, such as disability accommodations, illnesses, and grading problems. We review emails during ordinary business hours on weekdays, and may not reply in evenings or on weekends.

Study Partners The Student Academic Success Center will provide Study Partners sessions, which seek to create a study-group environment that helps students to become

independent, flexible problem-solvers. The sessions will be held for 2 hours weekly and open to any student to come and work on homework/problems together in a collaborative space. Attendance is not required from week-to-week. Students who attend should bring any questions, concerns, points of confusion, etc. to discuss with their classmates. Study Partners sessions will be facilitated by SASC student-staff members who are enrolled in the course. These facilitators will not assist with content beyond the skills of any other student enrolled in the course, but rather help organize and troubleshoot during this session.

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.

   - 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 ALL HOMEWORK PROBLEMS** (even if your lowest score is 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.

Research to Improve the Course

For this class, I am conducting research on student outcomes. This research will involve your work in this course. You will not be asked to do anything above and beyond the

normal learning activities and assignments that are part of this course. You are free not to participate in this research, and your participation will have no influence on your grade for this course or your academic career at CMU. If you do not wish to participate or if you are under 18 years of age, please send an email to Chad Hershock (hershock@andrew.cmu.edu) with your name and course number. Participants will not receive any compensation. The data collected as part of this research may include student grades. All analyses of data from participants coursework will be conducted after the course is over and final grades are submitted.

The Eberly Center may provide support on this research project regarding data analysis and interpretation. The Eberly Center for Teaching Excellence & Educational Innovation is located on the CMU–Pittsburgh Campus and its mission is to support the professional development of all CMU instructors regarding teaching and learning. To minimize the risk of breach of confidentiality, the Eberly Center will never have access to data from this course containing your personal identifiers. All data will be analyzed in de-identified form and presented in the aggregate, without any personal identifiers. If you have questions pertaining to your rights as a research participant, or to report concerns to this study, please contact Chad Hershock (hershock@andrew.cmu.edu).

Tentative Course Schedule[1]

(1) Week of 8/28	Course Mechanics. Motivation to Regression
	Quick Review of Random Variables
(2) Week of 9/4	Introduction to Simple Linear Regression
	Inference for β
(3) Week of 9/11	Estimating and Predicting with the Line
(4) Week of 9/18	Diagnosing and Correcting Problems
(5) Week of 9/25	Linear Regression in Matrix Notation
(6) Week of 10/2	Multiple Regression
	In-Class Exam One (Thursday, October 5)
	Data Exam One Out (Friday, October 6)
(7) Week of 10/9	Types of Predictors, Interactions
	Data Exam One Due (Friday, October 13)
(8) Week of 10/16	*Fall break*
	Fall break
(9) Week of 10/23	Anova
(10) Week of 10/30	Model Selection
(11) Week of 11/6	Effects of Predictor Relationships
	More on Effects of Predictor Relationships
(12) Week of 11/13	**In-Class Exam Two (Tuesday, November 14)**
	Data Exam Two Out (Tuesday, November 14)
	Weighted Least Squares
(13) Week of 11/20	Generalized Linear Models
	Data Exam Two Due (Tuesday, November 21)
	Thanksgiving break
(14) Week of 11/27	High-Dimensional Regression
(15) Week of 12/4	**In-Class Exam Three (Thursday, December 7)**

[1]**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.