36-401/601 Modern Regression

Course Policies and Syllabus, Fall 2025

Instructor: Zach Branson (Baker Hall 232H, zbranson@andrew.cmu.edu)
Office Hours: TBD

Teaching Assistants: Head TAs Woonyoung Chang (woonyouc@andrew.cmu.edu) and Xander Brick (abrick@andrew.cmu.edu), and a team of graduate/undergraduate TAs
Office hours: TBD

Lectures: Tuesdays/Thursdays 11:00am-12:20pm, Doherty Hall 2210

Course Website: On Canvas (https://canvas.cmu.edu/)
Lecture outlines, homeworks, and course announcements will be posted on Canvas. Make sure Canvas is configured to notify you when there are announcements about the class.

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

Textbook: No designated course textbook; you'll be provided all materials needed in lecture. Some relevant references (not required) are:

1. Sanford Weisberg, Applied Linear Regression, 4th Edition, 2013. (Available online through the CMU Library.)
2. Long & Teetor, *R Cookbook*, 2nd ed., O'Reilly, 2019. https://rc2e.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)
4. Wickham, ÃĞetinkaya-Rundel, and Grolemund. *R for Data Science*, 2nd ed., O'Reilly. https://r4ds.hadley.nz/

Course Overview

This course is an introduction to applied data analysis using linear regression modeling. We will derive properties about those models, apply and examine various models for real datasets, assess the validity of modeling assumptions, and determine what conclusions we can make (if any) from those models. We will use R to implement our analyses, produce graphs, and summarize 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, the questions of interest, and the appropriateness of different models to answer those questions with the data. Our overall goal is to use data and modeling tools to answer substantive questions, and to present and defend results in a scientific report.

The course includes a review of exploratory data analysis methods and statistical inference. We then consider simple linear regression models that only use one predictor variable. After briefly reviewing some linear algebra, we then consider multiple linear regression models that use multiple predictors. For all models, we will examine the appropriateness of their use for a given dataset (e.g., in terms of their implicit assumptions and inferential properties). Finally, we will explore extra topics such as, for example, generalized linear models and high-dimensional regression, time permitting. A tentative course schedule can be found on the last page of this syllabus.

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.

Because this course uses 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 techniques (e.g., graphical displays) to describe data, reveal pertinent phenomena for selecting statistical models, and aid in interpreting results from models to answer data-driven questions.
- Fit linear regression models based on given data and research questions, assess the validity of model assumptions, and interpret the results to provide accurate, data-driven conclusions.
- Understand and explain the mathematical theory underlying linear regression, and demonstrate deep understanding of how linear models are learned and interpreted.
- Develop comprehensive written arguments by using statistical evidence such as exploratory analysis, model fitting, and model evaluation, to answer data-driven questions. Be able to explain those findings to a general audience.
- Utilize R to implement effective data analyses.

Homework

There will typically be a homework each week, **due Fridays by 5:00pm**, except for around data exam times and when otherwise stated. All homeworks must be done in RMarkdown, and you must submit a PDF generated by RMarkdown to Gradescope. Gradescope will

be linked to the Canvas website so that this process is very straightforward. (For more information about Gradescope, see here.)

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.

Your homework submissions (compiled as a single PDF via RMarkdown and submitted to Gradescope) will include a mixture of textual answers, visualizations, code, and mathematical derivations. RMarkdown provides a comprehensive way to compile all of these materials as one PDF, which is why we require you to use RMarkdown to generate your homework submission. For mathematical derivations, we prefer that you do this with LaTeX within RMarkdown, but you may also write proofs by hand. If you prefer to write mathematical proofs by hand, you may scan your handwritten answers and merge your RMarkdown PDF with your scanned work to create a single PDF (note that Gradescope will only accept a single PDF). The aforementioned website about Gradescope also includes details about scanning handwritten work. You will be given guidance on how to use RMarkdown for the first homework.

It is your responsibility to read your PDF before submitting it, and make sure that it is readable for graders. If your PDF is not readable, we will have to deduct points accordingly. This applies not only to handwritten answers, but also to figures, code, and other parts of your submission. For example, if your code runs off the PDF as one long line—or worse, code is not provided—we will deduct points accordingly.

Furthermore, you must show your work and explain your reasoning for arriving at a particular answer. If only the correct answer is provided, but no relevant derivations, then zero points will be awarded. Similarly, if code is provided but it is not well-commented or organized (or textual answers are unnecessarily long and unorganized) such that it is difficult for us to understand what you are trying to implement and communicate, then your homework will be penalized accordingly. Finally, your work should be related to concepts and methods that we've discussed in class. If your work is far from concepts or methods that we've discussed in class (e.g., you use esoteric R packages that we have never discussed), then your homework will be penalized accordingly.

We will post homework solutions soon after the deadline. Graded homeworks will be available on Gradescope and Canvas, typically a week after the deadline. If you believe a mistake was made when your assignment was graded, **first check the assignment's solutions on Canvas**. If you still believe a mistake was made after—*and only after*—you have looked through the solutions, you can make a regrade request on Gradescope. Do not make a request via email. **Your regrade request must do two things; otherwise it will not be processed**:

- Explain how the solutions suggest that there may have been a grading mistake. **Be sure to refer to the solutions in your request.**

- The request must be made **within one week of when the assignment was graded**.

We grade according to a very detailed grading rubric; thus, regrade requests rarely result in points back.

Lastly, **you must write your homework on your own.** You can talk with others about homeworks (in fact, I encourage you to talk with your classmates about course content), but all work must be done by you. We will not accept homework copied from someone else (*any* "someone"—another student, some online entity, etc.). You are also not allowed to copy or "consult" solutions from previous semesters, or share course solutions with others, even after the end of the course. See the Academic Integrity section below.

Exams

There will be two in-person midterm exams, two take-home data analysis exams, and an in-person final exam. For all exams, you are not allowed to discuss the content of the exams with others until the solutions have been posted. **The dates of the exams will not be moved, so schedule job interviews and other extra-curricular activities around them.** The dates of the exams can be found in the schedule at the end of the syllabus.

In-Class Midterms: There will be two in-class midterms; the first midterm will focus on the first half-semester of material before the fall break, and the second will focus on the second half-semester of material after the fall break. The midterms will assess students' understanding of statistical theory/methodology as well as of data analysis. Although we will not ask you to handwrite code, we may ask you to communicate what a given set of code is doing, or interpret output/results from code and data analyses that are given to you. Exams will be designed to take approximately 40 minutes, but you will be given the full 80 minutes to complete the exam.

Final Exam: There will be an in-person final exam, to be given during finals week. The exam will be comprehensive, in the sense that it covers all material taught throughout the class. Similar to the midterms, the final exam will assess students' understanding of statistical theory/methodology as well as of data analysis.

Data Analysis Exams: There will be two take-home data analysis exams during the semester; students will be given a week to complete the exam, and there will be no homework during the exam. Data analysis exams will assess students' ability to implement various statistical models, interpret results from those models, use exploratory data analysis to evaluate models' appropriateness for a particular dataset, and justify modeling and data analysis choices to answer questions of interest. **During take-home exams, you *cannot* post on the class Piazza page; doing so will result in a zero for your exam.**

Further logistical details about each exam will be given closer to the exam dates. If you need special accommodations because of disabilities or religious holidays then please notify us as

soon as possible.

There will not be any makeup exams, so please plan accordingly. Again, see the last page of syllabus for exam dates.

Grading

Final grades will be computed according to the following weights:

Average Homework Score	20%
Lower In-Person Midterm Score	12.5%
Higher In-Person Midterm Score	17.5%
Lower Data Analysis Exam Score	12.5%
Higher Data Analysis Exam Score	17.5%
Final Exam Score	20%

No assignments are dropped when calculating midsemester grades. The lowest homework score will be dropped when calculating final grades.

Note that there are two in-person midterm exams. Among these two midterms, the lower midterm exam score will be given 12.5% weight, whereas the higher midterm exam score willl be given 17.5% weight. The two data analysis exams are similarly weighted. This systematically raises students' grades, compared to equal weighting.

Note that final grades are computed according to the weighting scheme above, and the "Total Grade" column on Canvas will not accurately reflect this weighting scheme. I do my best to make the Total Grade column on Canvas as accurate as possible, but there are (unfortunately) limitations to what Canvas can do. If you would like to know how you are doing in the class, simply compute your grade using the weighting scheme above. I always want the way you are being assessed to be as transparent as possible.

Final letter grades will be based on the standard 10-point scale: A = 90-100, B = 80-89, C = 70-79, D = 60-69, R = 0-59. I do round grades, such that an 89.50% would be an A, but an 89.49% would not. Final semester grades may be curved, but the curve will never lead to a worse grade than the 10-point scale. For example, a grade of 85% would earn you at least a B, and a curve could only improve it.

Lectures

Lectures are Tuesdays and Thursdays. Almost all course content and topics will be covered during lecture. Lecture notes will be posted on Canvas before each lecture; however, these notes will have many blanks that will be filled in with handwritten notes during lecture. Lecture notes will be organized by chapters; only *after* we have completed a chapter, will I

post handwritten notes on Canvas. It may take several classes for us to complete a particular chapter; thus, if you miss a given class, it may take one or two weeks for you to gain access to the handwritten notes. Furthermore, we will have a lot of discussion about the notes in class and tips for how to do well on the class assignments. **Thus, attending lecture is essential, because otherwise you'll miss a lot of material that will help you with assignments.**

Indeed, the primary purpose of lectures is to prepare you for class assignments (homeworks and exams). Thus, in addition to explaining course concepts, some lectures will involve practice problems that we'll work on together in-class. These problems will give you practice using course concepts, which in turn will help prepare you for the homeworks and exams. Throughout the lectures, you are expected to take notes and follow along with practice problems.

During lecture, I encourage you to ask questions about relevant course material we are discussing that day. Your questions show me what I have and have not made clear, which helps me teach more effectively. Furthermore, asking questions will help you become more comfortable with the material, and others will benefit from your perspective. If you have questions about the lecture material or practice problems after lecture is over, you should come to office hours.

Finally, no student may record, tape, or photograph any in-person classroom activity without the expressed written consent of the instructor. If a student believes that they have a disability and needs to record or tape classroom activities, they should contact the Office of Disability Resources to request an appropriate accommodation.

Communications, Email, and Piazza

If you have any questions about course content, please ask questions during class or come to office hours. Outside of class and office hours, the most reliable way to reach me and the Head TAs is by email. **However, Please use email only to address administrative and logistical issues.** For administrative and logistical issues, please email the Head TAs with the instructor CC'd, with "[36-401]" in the subject line. We will do our best to respond in a timely manner, but we will likely not respond in the late evenings, early mornings, or soon before an assignment deadline. It is always your responsibility to turn in assignments on time, regardless of whether or not we have replied to a particular email. If your email requires more than a short, non-technical response, we will likely say that we should talk in office hours instead.

There will also be a Piazza page for questions and discussion about class content. **When on Piazza (e.g., making a post, making a comment, etc.), you must adhere to the following rules:**

1. Be considerate to others (e.g., use respectful language, do not use sarcasm).

2. Before posting a question, check if it (or a related question) has already been posted. If it has, then use the existing thread for further questions/discussion. Do not create a new thread/post.
3. For questions that involve `R` code, questions like "What is wrong with this code?" are not acceptable. Code that is part of your solution cannot be posted to Piazza.
4. Along with your posted question, explain what you tried to answer your own question (without posting solutions).

Content deemed inappropriate—by the above rules or otherwise—will be taken down by the instructor or TAs. Furthermore, **you cannot post on Piazza during the data analysis exams (discussed above); doing so will result in a zero for your exam.**

The primary purpose of the Piazza page will be to ask quick clarifying questions. Similar to email, if your Piazza post requires a detailed and/or technical response, we will ask you to instead attend office hours.

We will regularly check Piazza to answer questions, but please allow for up to 48 hours for us to answer your questions. **This means that posting a question on Piazza the day the homework is due will likely go unanswered.** Even if you make a post on Piazza, you are still responsible for the on-time, high quality completion of assignments.

Given the above, attending office hours will often be a quicker way for you to get your questions answered effectively.

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

Although you are encouraged to discuss course material with classmates (except during the take-home exams), copying answers (even incorrectly) from someone else, allowing someone else to copy from you, or any other form of cheating are typically grounds for course failure. The previous sentence applies even if "someone else" is an online entity. In these situations, it is my duty to report the incident to the appropriate University authorities. This applies to all assignments.

You may use external resources to help with your work, but they must *assist* your own thinking, and 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 (or other generative AI tools) to do assignments for you. If you're ever unsure how this academic integrity policy applies to this class, please reach out to me via email so that we can discuss.

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 or otherwise sharing them with future students is also an academic integrity violation, and is considered "unauthorized assistance" under CMU's policies.

Finally, it is worth noting that course assignments (e.g., the homeworks) are designed to help you prepare for future course assignments (e.g., the exams). Thus, if you do not put meaningful effort into the course assignments (e.g., by using generative AI tools to do parts of the assignment for you), then you will likely not learn from the course assignments. In this case, it is much more likely that you will not do well on the exams. This is a common reason why students have failed the course in the past.

Accommodations for Students with Disabilities

If you require a special accommodation (e.g., for exams, class notes, or other aspects of the class), please visit the Office of Disability Resources here to obtain appropriate documentation. I cannot make any accommodations without documentation.

Study Tips

I genuinely want everyone to do well in this class—below are some tips for how to do so.

1. **Attend class and actively take notes**. The professor will 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.
    - 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.

- For steps in the notes where you are confused, try to fill in gaps with additional explanations.
- Write down questions about things you do not understand. Bring these questions to us during office hours 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, online resources, or any other external resources. Otherwise, you run the risk of not effectively learning from the homeworks, thereby increasing the chance that you will not do well on the exams.
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.

A Note on Diversity

My aim is that this course is well-served by all backgrounds and perspectives, and I want all students' learning needs be addressed both in and out of class. One of the best things about Carnegie Mellon is that an amazingly diverse group of people come together to learn, work, and achieve their goals. I value that diversity deeply—it is a resource, strength, and benefit of our community. It is my intent that class materials and activities are respectful of this diversity, whether it be in terms of gender, sexuality, disability, age, socioeconomic status, ethnicity, race, and/or culture. Your suggestions are encouraged and appreciated. Please let me know if there are ways that I can improve the effectiveness of the course for you personally or for other students. That said, I will continually work to make this course as welcoming and effective as possible, rather than just wait for students to tell me what's not working. More generally, I'm always open to hearing your thoughts on how statistics can be more welcoming to people from any background.

A Note on Life-Work Balance

I care deeply about this course material and am excited to teach it, but I care even more so about your well-being. I did my undergraduate at Carnegie Mellon (2010-2014), and I had to learn the hard way that physical and mental health are much more important than grades—I don't remember my CMU grades, but I do remember my lifelong friends and mentors, as well as the many fun memories I made in Pittsburgh. During my undergrad years, the university

was just starting to have a conversation about stress culture, so that issue is very important to me. I am always happy to talk with you about life at CMU and Pittsburgh, your future career and education, or just about anything, really, if it supports your well-being here. There are many resources around you—family, friends, advisors, mentors, peer tutors (which I was at CMU!), RAs (which I was also at CMU!), housefellows, etc.—and you should take advantage of them. I also encourage you to be aware of professional resources:

- Counseling and Psychological Services (CaPS; 412-268-2922 or http://www.cmu.edu/counseling/).
- If you are worried about affording food or feeling insecure about food, email the CMU Food Pantry Coordinator to schedule an appointment (cmu-pantry@andrew.cmu.edu, 412-268-8704).
- If you or someone you know is in a life-threatening situation, call the police immediately (8-2323 on campus, 911 off campus).

I genuinely want your experience in this class and at CMU to be rewarding and enjoyable. I do my best to make this class engaging, accommodating, and professionally valuable, but I never want the time you put into this class to be at the detriment of your well-being. Please reach out whenever you have concerns about the course, and/or ways I can help improve your experience in this class and at CMU.

Tentative Course Schedule

Week 1 (Aug. 26 & 28)	Course Overview and Motivation Review of Random Variables, Probability Theory, Inference
Week 2 (Sept. 2 & 4)	Simple Linear Regression, Inference for β Homework1 due Fri. Sept. 5 by 5pm
Week 3 (Sept. 9 & 11)	Estimation and Prediction, Bias and Variance Homework2 due Fri. Sept. 12 by 5pm
Week 4 (Sept. 16 & 18)	Assumptions, Diagnostics, Corrections Homework3 due Fri. Sept. 19 by 5pm
Week 5 (Sept. 23 & 25)	Matrix Notation, Multivariate Random Variables Homework4 due Fri. Sept. 26 by 5pm
Week 6 (Sept. 30 & Oct. 2)	Multiple Regression **In-Class Midterm1 (Thursday, October 2)** *Data Analysis Exam1 Released (Friday, October 3)*
Week 7 (Oct. 7 & 9)	Types of Predictors, Interactions **Data Analysis Exam1 Due (Friday, October 10 by 5pm)**
	FALL BREAK (Oct. 13-17)
Week 8 (Oct. 21 & 23)	ANOVA, Model Selection
Week 9 (Oct. 28 & 30)	Model Selection, Cross Validation Homework5 due Fri. Oct. 31 by 5pm
Week 10 (Nov. 4 & 6)	Effects of Predictor Relationships NO CLASS TUESDAY, NOV. 4 (Democracy Day) Homework6 due Fri. Nov. 7 by 5pm
Week 11 (Nov. 11 & 13)	Generalized Linear Models Homework7 due Fri. Nov. 14 by 5pm
Week 12 (Nov. 18 & 20)	High-Dimensional Regression **In-Class Midterm2 (Tuesday, November 18)** *Data Analysis Exam2 Released (Tuesday, November 18)*
Week 13 (Nov. 25 & 27)	Special Topics (Time-permitting) **Data Analysis Exam2 Due (Tuesday, November 25 by 5pm)** NO CLASS THURSDAY, NOV. 27 (Thanksgiving)
Week 14 (Dec. 2 & 4)	Special Topics (Time-permitting) **Final Exam during Finals Week (TBD)**