

Course Syllabus

 Edit

IDS/Ec/PS 126: Applied Data Analysis

(<https://caltech.instructure.com/courses/5979/pages/home-page>)

Fall 2023

Course Instructor

Prof. Jonathan N. Katz (he/him)  (<https://jkatz.caltech.edu/>)

 jkatz@caltech.edu (<mailto:jkatz@caltech.edu>)

 [@Jonathan_N_Katz](https://twitter.com/jonathan_n_katz)  (https://twitter.com/jonathan_n_katz)

 **Office hours: Sign up here**  (<https://calendly.com/jnkatz/student-meeting>)


Teaching Assistant

TBD  (<http://www.hss.caltech.edu/people/shiyu-zhang>)

Course Details

 **Mondays and Wednesdays**

 **September 27 - November 29**

 **10:30am - 12:00pm PST**

 **Baxter 125**

Course Description

Fundamentally, this course is about making arguments with numbers and data. Data analysis for its own sake is often quite boring, but becomes crucial when it supports claims about the world. A convincing data analysis starts with the collection and cleaning of data, a thoughtful and reproducible statistical analysis of it, and the graphical presentation of the results. This course will provide students with the necessary practical skills, chiefly revolving around statistical computing, to conduct their own data analysis. This course is not an introduction to statistics or computer science. I assume that students are familiar with at least basic probability and statistical concepts up to and including regression.

Course Philosophy

Most standard statistics classes typically focus only on one small part of the process of generating a persuasive analysis of data, the mathematical and probability theory underlying various methods. In practice, a significant amount of time is spent finding or creating the data

and then cleaning and verifying it. Then the process of statistical model building is iterative: a model is fit, evaluated, and then possibly refined several times. Finally, once the final model is chosen, we need to communicate it to our audience. This course will focus on this entire chain including practical advice that is often omitted from standard courses.


Learning Outcomes

By the end of this course you will be able to:

- Understand the fundamental types of data analysis: exploratory, predictive, and causal inference.
- Build and critically evaluate regression models to use in applied analyses.
- Produce reproducible analyses workflows.
- Graphically present results of a data analysis.
- Wrangle, clear, and transform data.
- Write, execute and debug R code.

Required Texts

The main textbook for the course is:



- Andrew Gelman, Jennifer Hill and Aki Vehtari. [Regression and Other Stories](https://doi.org/10.1017/9781139161879)  (<https://doi.org/10.1017/9781139161879>). Cambridge University Press. 2020.


We will also have some readings from the open source (free):

- Garrett Golemund and Hadley Wickham. [R for Data Science](https://r4ds.had.co.nz)  (<https://r4ds.had.co.nz>).

Course Software

The central home for this course is in [Canvas \(https://caltech.instructure.com/courses/5979/pages/home-page\)](https://caltech.instructure.com/courses/5979/pages/home-page). For discussion forums we will be using Piazza. These can both be accessed from within Canvas on the navigation bar to the left.

You will do all of your statistical analysis with the open source programming language [R](https://cran.r-project.org)  (<https://cran.r-project.org>). You will use [RStudio](https://www.rstudio.com/)  (<https://www.rstudio.com/>) as the main program to access R. Formally, RStudio is an integrated development environment (IDE). Think of R as an engine and RStudio as a car dashboard—R handles all the calculations and the actual statistics, while RStudio provides a nice interface for running R code. Formally, RStudio is an integrated development environment.

R is free, but it can sometimes be a pain to install and configure. To make life easier, I have arranged for the use [RStudio Cloud](http://rstudio.cloud/)  (<http://rstudio.cloud/>) service for the course, which lets you run a full instance of RStudio in your web browser. This means you won't have to install anything on your computer to get started with R! We will have a shared class workspace in RStudio.cloud that will let you quickly copy templates for labs and problem sets.

RStudio Cloud is convenient, but it can be slow and it is not designed to be able to handle larger datasets or more complicated analysis. You may want to install R, RStudio, and other R packages on your computer to work on your final project. This is not necessary, but it may be helpful.

Attendance and Participation

Although I strongly encourage attendance and participation, given the situation around COVID, they will not affect your grade. **If you are not feeling well, please do not attend class.** I will be covering some material that is not the course readings during the lectures. All course sessions will be reordered and will be available on the course website.

Academic Integrity

This course is governed by Caltech's Honor Code: *"No member of the Caltech community shall take unfair advantage of any other member of the Caltech community."*

Understanding and Avoiding Plagiarism: Plagiarism is the appropriation of another person's ideas, processes, results, or words without giving appropriate credit, and it violates the honor code in a fundamental way. You can find more information at: <http://writing.caltech.edu/resources/plagiarism>  (<http://writing.caltech.edu/resources/plagiarism>).

Collaboration Policy

Collaboration on homework assignments is encouraged. You may consult outside reference materials, other students, or the instructor, but you cannot consult homework solutions from prior years and you must cite any use of material from outside references. You may consult generative AI tools to help, but you can not copy code verbatim. All solutions that are handed in should be written up individually and should reflect your own understanding of the subject matter at the time of writing. Your code are considered part of your write-up and should be done individually. You can share ideas, but not code. As already noted, the final project is an individual

Grading

Your final grade in the course will be composed of two components with the following weights:

Problem Sets	70%
Final Project	30%


Participation: Although I strongly encourage attendance and participation, given the situation around COVID, they will not affect your grade. Again if you are not feeling well, please do not attend class.


Problem Sets: There will be eight weekly problem that are **due the following Tuesday evening at 9:00pm PST**. They will also be available via the weekly Canvas modules. In general, I do not accept late assignments. However, given the unusual world circumstances, I will entertain accepting late assignment on a case-by-case basis. I will be much more likely to grant an extension to a particular assignment if you contact me prior to the due date for it.

Final Project: You will demonstrate your knowledge of applied data analysis by completing your own original project. There will be various interim components due during the term and the final project write-up and replication materials are due on **Friday, December 10 at 5:00pm PST**.

Schedule

- **[Week 1 \(September 27\): Course Overview \(https://caltech.instructure.com/courses/5979/pages/week-1-class-review-and-todo\)](https://caltech.instructure.com/courses/5979/pages/week-1-class-review-and-todo)**
 - Readings:
Regression and Other Stories, ch. 1

- **[Week 2 \(October 2\): R, Tidyverse, and RMarkdown \(https://caltech.instructure.com/courses/5979/modules/41358\)](https://caltech.instructure.com/courses/5979/modules/41358)**
 - NOTE: **There will not be class on Wednesday, October 5**
 - Problem set due **[Tuesday, October 10 at 9:00pm PST \(https://caltech.instructure.com/courses/5979/assignments/47461\)](https://caltech.instructure.com/courses/5979/assignments/47461)**
 - Readings:
[R for Data Science, ch. 4, 5, 10, and 27](https://r4ds.had.co.nz)  (<https://r4ds.had.co.nz>)

- **[Week 3 \(October 9\): Data and Measurement \(https://caltech.instructure.com/courses/5979/modules/41359\)](https://caltech.instructure.com/courses/5979/modules/41359)**
 - Problem set due **[Tuesday, October 17 at 9:00pm PST \(https://caltech.instructure.com/courses/5979/assignments/47462\)](https://caltech.instructure.com/courses/5979/assignments/47462)**
 - Readings:
[R for Data Science, ch. 3 and 7](https://r4ds.had.co.nz)  (<https://r4ds.had.co.nz>)
Regression and Other Stories, ch. 2

- **[Week 4 \(October 16\): An Unconventional Take on Statistical Inference and Regression Basics \(https://caltech.instructure.com/courses/5979/modules/41360\)](https://caltech.instructure.com/courses/5979/modules/41360)**
 - Problem set due **[Tuesday, October 24 at 9:00pm PST \(https://caltech.instructure.com/courses/5979/assignments/47463\)](https://caltech.instructure.com/courses/5979/assignments/47463)**

- [/courses/5979/assignments/47463](#)
 - Readings:
Regression and Other Stories, ch. 4, 6, and 7

- [Week 5 \(October 23\): Fitting Regression models and predictions](#)
[\(https://caltech.instructure.com/courses/5979/modules/41361\)](https://caltech.instructure.com/courses/5979/modules/41361)
 - Final Project Proposals due [Friday, October 27 at 9:00pm PST](#)
[\(%24CANVAS_OBJECT_REFERENCE%24/assignments/g0bda138c223075341d71b1e7e96c3a66\)](https://caltech.instructure.com/courses/5979/assignments/g0bda138c223075341d71b1e7e96c3a66)
 - Problem set due [Tuesday, October 31 at 9:00pm PST](#) (<https://caltech.instructure.com/courses/5979/assignments/47464>)
 - Readings:
Regression and Other Stories, ch. 8, 9, and 10

- [Week 6 \(October 30\): Model Evaluation and Transformations](#)
[\(https://caltech.instructure.com/courses/5979/modules/41362\)](https://caltech.instructure.com/courses/5979/modules/41362)
 - Problem set due [Tuesday, November 7 at 9:00pm PST](#) (<https://caltech.instructure.com/courses/5979/assignments/47465>)
 - Readings:
Regression and Other Stories, ch. 11 and 12

- [Week 7 \(November 6\): Generalized Linear Models](#) (<https://caltech.instructure.com/courses/5979/modules/41363>)
 - Problem set due [Tuesday, November 14 at 9:00pm PST](#) (<https://caltech.instructure.com/courses/5979/assignments/47466>)
 - Readings:
Regression and Other Stories, ch. 13 and 14

- [Week 8 \(November 13\): Design and Poststratification](#) (<https://caltech.instructure.com/courses/5979/modules/41364>)
 - Final Project Interim Results due [Wednesday, November 22 at 9:00p PST](#)
[\(https://caltech.instructure.com/courses/5979/assignments/47459\)](https://caltech.instructure.com/courses/5979/assignments/47459)
 - Problem set due [Tuesday, November 21 at 9:00pm PST](#) (<https://caltech.instructure.com/courses/5979/assignments/47467>)
 - Readings:
Regression and Other Stories, ch. 16 and 17

- [Week 9 \(November 20\): Causal Inference I](#) (<https://caltech.instructure.com/courses/5979/modules/41365>)

- Problem set due **Tuesday, November 28 at 9:00pm PST** (<https://caltech.instructure.com/courses/5979/assignments/47468>)
- Readings:
Regression and Other Stories, ch. 18 and 19

- **Week 10 (November 27): Causal Inference II** (<https://caltech.instructure.com/courses/5979/modules/41366>)
 - Readings:
Regression and Other Stories, ch. 20 and 21

- **Week 11 (December 4):**
 - **No Class, but final project due by Friday, December 8 at 5:00pm PST** (<https://caltech.instructure.com/courses/5979/assignments/47458>).