432 Spring 2020 Syllabus

Thomas E. Love, Ph.D.

Version 2020-01-14 08:47:30

Contents


CASE
WESTERN
RESERVE
EST. 1826 UNIVERSITY

Key Information

This is the Spring 2020 syllabus page for PQHS / CRSP / MPHP 432: Data Science for Biological, Medical and Health Research II, taught by Professor Thomas E. Love. The course is given on Tuesdays and Thursdays from 1:00 to 2:15 PM, in Room E321-323 in the Robbins building of the CWRU School of Medicine.

Course Home Page

The course home page, with links to everything you'll need, is at https://github.com/THOMASELOVE/2020-432.

- All class meetings, deadlines and assignments are listed in the Course Calendar.

Getting Help!

To get help for anything related to the course, email the Teaching Assistants and Dr. Love at `431 dot help at case dot edu`.

- Dr. Love is available on Tuesdays and Thursdays at CWRU, by appointment. To make an appointment, email him directly at `thomas dot love at case dot edu`. His office is Wood WG-82 J.
- If you have any special concerns about the course, need special accommodations or any other issues for Dr. Love, please email him, or speak with him before or after class.

Chapter 1

Course Description

PQHS 432 (cross-listed as CRSP 432 and MPHP 432, and formerly known as EPBI 432) is the second half of a two-semester sequence (with PQHS 431) focused on modern data analysis and advanced statistical modeling, with a practical bent (as little theory as possible), emphasizing the key role of thinking hard, and well, about design and analysis in research. The title listed by the registrar is a little dated - I prefer *Data Science for Biological, Medical or Health Research*.

This is a good course for people who want to learn how to use the R language to get information from data, and who want to learn about making comparisons and building models to help make meaningful progress in research, focusing on questions from biology, medicine and public health. We spend time managing and visualizing data, building models and making predictions, and other things thought of as "data science" - in essence, this highly applied course focuses on modern, more than classical, tools for learning from data. The course is taught using the R statistical software and RStudio environments, with the material discussed in 431 assumed in 432. Students learned a lot of R in the 431 course, and that material remains available at https://github.com/THOMASELOVE/2019-431. We'll continue to use R Studio and R Markdown as tools to help make R work better, and perform our research in replicable ways.

1.1 General Approach / Topics

The course covers the following general topics, roughly in this order, through early April. Additional topics (for the remainder of April) will be determined later in the semester.

1. Linear Regression (including weighted and robust approaches, variable selection, dealing with missing data, fitting non-linear relationships through

predictor transformation, cross-validation approaches, and multi-factor ANOVA and ANCOVA)
2. Logistic Regression (including both models for binary outcomes, and models for proportions, and risk adjustment)
3. Generalized Linear Models (including regression models for count data, multi-categorical outcomes)
4. The Statistical Crisis in Science
5. Cluster Analysis (mostly in the form of Principal Components Analysis)
6. Survival Analysis (Kaplan-Meier curves and Cox Regression)

1.2 Prerequisites

Taking 432 without 431 is not recommended. The pace can be brisk at times, but all CWRU students who feel up to it are welcome, in any field of study.

The main things students need for 432 are:

- tools: substantive knowledge of the use of R, R Studio and R Markdown to produce code which will ingest, visualize, explore, analyze and model data, then communicate the results
- statistical methodology: substantive understanding of statistical inference in the one-, two- and multi-sample cases and the fundamentals of linear regression models, including the building of multiple linear regressions, and their evaluation through diagnostic plots, stepwise model selection, assessment of uncertainty via confidence and prediction intervals, and basic in-sample and out-of sample validation summaries
- data to study related to biological, health, medical, scientific or other phenomena, and
- an interest in studying data closely and presenting rigorous analyses effectively

Some of these topics are reviewed in early 432 sessions.

1.3 Everything is on the Web

https://github.com/THOMASELOVE/2020-432 is the place to go for everything related to this course. Please visit any time you need something. I update the web site frequently.

- The most important thing is the Course Calendar which serves as the final word for all deadlines, plus links to all classes and deliverables.
- Dr. Love's book of 432 Course Notes is the principal textbook for the course, and will appear during the semester. The Spring 2019 version is available now.

Additional details will be coming soon.

Chapter 2

Professor Love

Thomas E. Love, Ph.D.

- Professor of Medicine, Population and Quantitative Health Sciences, CWRU
- Director of Biostatistics and Evaluation, Center for Health Care Research & Policy, MetroHealth Medical Center
- Chief Data Scientist, Better Health Partnership
- Track Lead for Health Care Analytics, MS in Biostatistics, Department of Population and Quantitative Health Sciences, CWRU
- Fellow, American Statistical Association

2.1 A More Complete Biography

Hi. I am Thomas E. Love, Ph.D. and I have at least three different jobs.

- I am a Professor in the Departments of Medicine and Population & Quantitative Health Sciences at Case Western Reserve University. I teach three

courses per year there (PQHS 431, 432 and 500) and also lead the Health Care Analytics track of the MS program in Biostatistics.
- I direct Biostatistics and Evaluation at the Center for Health Care Research & Policy, which is a joint venture of CWRU and MetroHealth Medical Center.
- For ten years, I was the (founding) Data Director for Better Health Partnership, an alliance of people who provide, pay for and receive care in Northeast Ohio. I now serve as Chief Data Scientist there.
- I am a Fellow of the American Statistical Association, and have won numerous awards for my teaching and my research, including the 2018 John S. Diekhoff Award for Graduate Teaching from CWRU.
- I have been teaching at CWRU since 1994, and have taught every manner of CWRU student over the years, especially students in biostatistics, medicine, and management.

In research, I use statistical methods to look at questions in health policy and in particular the provision of health services. I mostly work with observational data, rather than data that emerge from randomized clinical trials, and I have a special interest in working with data from electronic health records.

- You may be interested in a recent study in Health Affairs showing the impact of a Medicaid-like expansion plan on care and outcomes of poor patients in Cleveland.
- Or you might be interested in our New England Journal of Medicine study of the effect of electronic health records on the care and outcomes of people with diabetes.
- In 2011, James O'Malley and I chaired the Ninth International Conference on Health Policy Statistics, here in Cleveland. Here's a recap.
- I've also worked on many projects involving the use of propensity scores to make causal inferences from observational studies, particularly in heart failure.

If you want to see a pretty complete list of my publications, knock yourself out.

I hold degrees from Columbia University in the City of New York and from the University of Pennsylvania. My dissertation adviser was Paul Rosenbaum. I am married to a brilliant woman who is an attorney at GE Lighting, and we are raising two terrific sons. The elder is a junior in college (University of Pittsburgh) who plans to be a paleontologist, and one finishing high school this year, who will attend Columbia this Fall. I live in Shaker Heights. I also sing and act occasionally in community theater.

2.2 Email

- Email to get help with the course: **431-help at case dot edu** (seen by Professor Love and the TAs)

- `Thomas dot Love at case dot edu` (for matters related to grades or individual concerns)
- Professor Love is hard to reach by phone. Email is always the best way to reach him.

2.3 Offices

- Wood WG-82J on the ground floor of the Wood building (Tuesdays and Thursdays)
- Rammelkamp R-229A at MetroHealth Medical Center (Wednesdays and Fridays)

Professor Love is available for the 15 minutes before and the 30 minutes after each class, and otherwise by appointment on Tuesdays and Thursdays (send email to schedule).

2.4 Name and Pronouns

- Professor Love uses he/him/his pronouns.
- Most students refer to him either as Professor Love or Dr. Love.
- He prefers his given name to be written "Thomas" as opposed to "Tom".
- Most of his friends and colleagues call him "Tom". You are welcome to do so, as well, if that makes you more comfortable.

2.5 Web

- Professor Love's GitHub pages website.
  - His GitHub name is THOMASELOVE.
- His Twitter handle is ThomasELove.

Chapter 3

Teaching Assistants

- To contact the TAs (and Dr. Love) at any time, email `431-help` at `case dot edu`.

The teaching assistants for 432 this year are

- Benjamin (Ben) Booker, BS
- Julijana Conic, MD
- Joseph Hnath, BA
- Amr Mahran, MD MS
- Amin Saad, MD
- Jing Zhang, MD MS

They are the people answering `431-help at case dot edu`, and they are the people holding the bulk of our regular office hours. Most of them has been in your shoes - they've taken the course in the past, and they enjoyed it enough to come back for more. Many have volunteered their precious time and energy to help make the course happen, and we couldn't be more delighted to welcome you to the course. To contact the TAs, email `431-help at case dot edu`, which is open all semester, starting on the first day we meet.

3.1 Office Hours for TAs

- To contact the TAs (and Dr. Love) at any time, email `431-help` at `case dot edu`. This is a challenging class. Don't suffer in silence - talk to us!

Teaching Assistant Office Hours are held in either WG-56 (Computing Lab) or WG-67 (Student Lounge) on the ground floor of the Wood building in the School of Medicine, so be sure to look in both places if you need help. The weekly schedule will be posted on the bottom of the Course Calendar.

TA office hours are not held on University holidays, or during Spring Break, although 431-help remains open until the last project is completed in May.

This is a challenging class. Don't suffer in silence - talk to us!

3.2 Benjamin (Ben) Booker, BS

Benjamin (Ben) Booker is a first year PhD student in the Epidemiology & Biostatistics program in the Department of Population & Quantitative Health Sciences. Ben holds a BS in Molecular Biology from the University of Cincinnati, and then completed two years of additional training in Biostatistics there. He has worked at Cincinnati Children's Hospital performing DNA methylation analysis, and as a data scientist consultant for Givaudan Flavors. Outside of work and school I enjoy rock climbing/bouldering (novice level), playing soccer and watching the European football leagues.

3.3 Julijana Conic, MD

Julijana Conic was born in Serbia and received her MD from the University of Belgrade Faculty of Medicine last year. Since enrolling in the MS in Clinical Research program the same year she has been conducting research focusing on ischemic mitral regurgitation in the Department of Cardiovascular Imagining at the Cleveland Clinic. Currently, she is working on a project incorporating

machine learning to improve existing algorithms for automatic quantification of cardiac volumes on MRI images and to aid in risk stratification of ischemic mitral regurgitation patients. She hopes to start internal medicine residency next year and ultimately establish herself as a physician investigator. During her free time Julijana enjoys hiking, watching movies and volunteering in the community.

3.4 Joseph Hnath, BA

Joseph Hnath is in his second year of the Master of Public Health program on the Intensive Research Pathway with concentrations in Population Health Research and Health Policy & Management. He finished his undergraduate studies at CWRU this May where he majored in Chemical Biology, Cognitive Science, and Economics. Having taken 431 & 432 last year, the skills he learned have been invaluable in his research projects, such as his capstone on the health economics of abortion policy and helping with the NEO-CASE cancer disparities resource. Joseph enjoys playing basketball, watching Master Chef, and reading *The Complete Works of F. Scott Fitzgerald.*

3.5 Amr Mahran, MD MS

Amr Mahran is a urologist who is working as a senior research associate in the department of urology, CWRU School of Medicine. He received his MD degree from Assiut University School of Medicine in Upper Egypt. He also finished a residency in urology along with earning a Master of Science degree. Before joining CWRU, Amr was a practicing urologist and was appointed as a faculty at the department of urology, Assiut University Hospitals. Amr took 432 in the spring of 2019 and learned many skills that helped him in his clinical research. Amr's research focus on prostate cancer, pelvic pain, and voiding dysfunction. He does outcome research on large databases as NSQIP, National Trauma Database (NTDB), and NIS databases. Amr enjoys playing soccer, table tennis, and reading.

3.6 Amin Saad, MD

Amin Saad is an international medical graduate from Syria with two years of General Surgery residency training experience in the United States. Amin is

currently enrolled in the CRSP Master's program and is seeking a Ph.D. degree in Clinical and Translational Research with a focus toward lowering surgical site infection rates. Amin took 431 and 432 two cycles ago and has appreciated how the skills he learned in those classes have helped him with his clinical outcomes research at the Department of Colorectal Surgery at University Hospitals. Amin enjoys playing soccer, swimming, and spending time with his family.

3.7 Jing Zhang, MD MS

Jing Zhang is a first year PhD student in the Epidemiology and Biostatistics program in the Department of Population and Quantitative Health Sciences. Jing finished her undergraduate studies in Clinical Medicine and graduate studies in Biostatistics at Fudan University, Shanghai, China. She values the statistical analysis skills learned during graduate studies and enjoys solving statistical problems. During her spare time, she likes jogging and cooking.

Chapter 4

Deliverables and Grading

4.1 Timing and Deadlines

The Calendar is the exclusive home for all deadlines in the course.

4.2 Participation in the Course

Students are required to participate actively in the course, including meaningful contributions in group work, in-class and minute paper participation, emails to 431-help, visits to the TAs, etc.

- We're more concerned about the breadth of your participation rather than just its quantity.
- If you're having trouble asking questions, the best way to make a contribution is to find something interesting and share it with us, through 431-help.
- Most students score between 80% and 100% on this element.

4.2.1 Attendance

I expect you to come to class. If you have to miss a single class, just be sure to catch up on any needed materials - no need to let me know in advance or afterwards. We expect you to complete all necessary deliverables, and to review the README for that day's class for other announcements. The audio recording can help, too.

If, however, you are going to miss more than one class in a row, you should let Dr. Love know, via email, in advance, ideally.

4.3 Projects

Students are required to complete two project assignments, one in mid-semester, and one at the end of the term.

Details on the project assignments are posted on the Course Projects Page.

4.4 Homeworks

There will be 6-8 homework assignments this semester. The exact number is not settled yet.

Details on those assignments are posted on the Homework Page.

4.5 Quizzes

Students will complete several quizzes during the semester. The exact number is not settled yet.

Details on the Quizzes are posted on the Course Quizzes Page.

4.6 Grading

The final course grade is weighted as follows:

- 15% Class Participation
- 20-30% Homework
- 15-25% Quizzes
- 40% Two Projects, including the Final Portfolio Presentation

A cut point to discriminate A vs. B will be set in the range of 85% to 90% at the end of the term. An average of 70% or higher is required to receive a B. Final decisions on the relative weights of the Homework and Quizzes will depend primarily on the number of Homeworks and number of Quizzes assigned.

Chapter 5

A Few Writing/Presenting Tips

1. Statistics is a "getting the details right" business - we care deeply about details, and this applies to writing code or complete English sentences.
2. Nothing impresses us as much as a clear and concise argument, presented using well-written English sentences, effective and well-labeled figures and tables.
3. Don't parrot back material that Dr. Love wrote or said. State ideas in your own words. Stating them in other words is, technically, plagiarism.
4. Edit your more adventurous output; don't present everything you know how to do in R, and don't forget that someone is trying to read both your code and your results.
5. Make your work easy to evaluate. In responding to an assignment, be sure to answer the question that was asked, restating it as necessary.
6. Clearly label everything: graphs, tables, your answer to a specific question. Everything. Again, make your work easy to evaluate.
7. Simplify. Emphasize ideas in plain language. Avoid jargon. Use English well.
8. Data are plural. Use "the data **are** ... " rather than "the data *is* ... "
9. A paragraph must contain more than one sentence.
10. Don't switch tenses. If you want to write in the present tense, stick to it throughout.
11. Don't write or say random sample unless you used a random number generator. If you used haphazard sampling or convenience sampling, call

it what it is, and indicate whether any problems could have cropped up as a result.

12. Similarly, don't defend a method of data collection because it is random. Most of the time we want to represent some population, and a random sample is just one way to ensure that certain types of biases have a low probability of creeping in.

13. If you want to write that you used $\alpha = 0.05$ as your significance level, then state that your results were obtained using a 95% confidence level, not a 95% confidence interval, unless you are actually interpreting a confidence interval.

14. If you're looking at a p-value, then you should state either:

    - [1] We're using a 95% confidence level.
    - [2] We're using a 5% significance level. or
    - [3] We're using $\alpha = 0.05$.
    - Don't use more than one of these expressions.

15. Refer to all p-values that are less than 0.001 or perhaps less than 0.0001 as $p < 0.001$, rather than, for instance, $p = 0.00000001$ or, worse yet, $p = 0$. In a similar vein, write all p-values that exceed 0.99 as $p > 0.99$ instead of, for instance, $p = 1$.

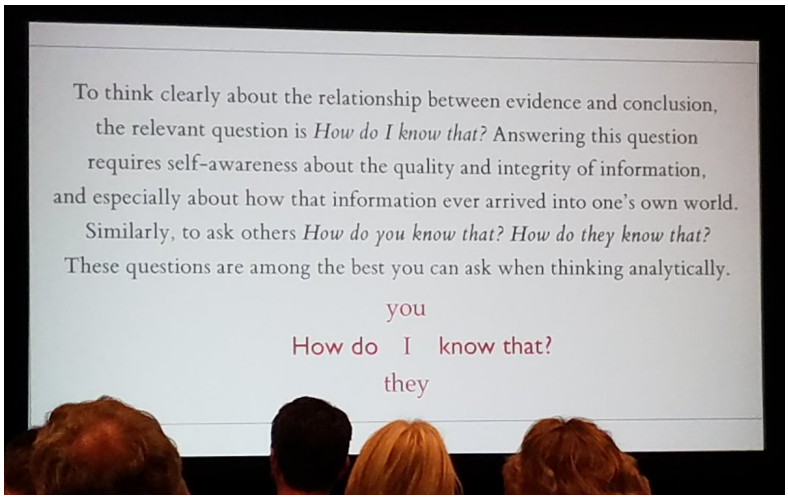
16. To the extent possible, don't use `computer-ese` to label variables, plots or tables. R and Markdown allow you to change the labels on graphs and tables to meaningful things – do so. Use meaningful abbreviations, as necessary, explaining what they mean on the first usage.

17. Use words that we all know, whenever possible, and provide clear definitions at the first encounter when jargon is mandatory.

18. Often the most useful thing you can do in an analysis is to turn a table into a meaningful graph.

19. When in doubt, err on the side of clearer expression. Clear thinking causes and is demonstrated by clear writing.

20. In the words of Edward Tufte, to think clearly, keep asking yourself ...


To think clearly about the relationship between evidence and conclusion, the relevant question is *How do I know that?* Answering this question requires self-awareness about the quality and integrity of information, and especially about how that information ever arrived into one's own world. Similarly, to ask others *How do you know that? How do they know that?* These questions are among the best you can ask when thinking analytically.
you
How do I know that?
they