432 Spring 2019 Syllabus

Thomas E. Love, Ph.D.

Version 2019-01-21 22:36:57

Contents

Key Information

This is the Spring 2019 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/2019-432.

Before Class 1

Visit the course home page for a list of things you need to do before our first class meeting on Tuesday **2019-01-22**. Note that Dr. Love is away the first week of the semester.

- All class meetings 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 L.
- 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, for example, 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/431-2018. 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: substantiave 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 and/or medical 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/2019-432 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 Course Calendar serves as the final word for all deadlines, plus links to all classes and deliverables.
- Dr. Love's book of 432 Course Notes is now available. This is the principal textbook for the course, and will be updated occasionally during the semester.
- Data and Code for the course are now available.
- Readings and Supplemental Materials appear on the References page.
- The Slides page provides links to presentation materials and READMEs for each class session.

Also available on the course website are instructions and information on the course deliverables, including projects, homework and quizzes.

Chapter 2

Dr. 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 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)
- Dr. Love is hard to reach by phone. Email is always the best way to reach him.

2.2 Offices

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

Dr. Love is generally available for a few minutes before and 30 minutes after class, otherwise by appointment on Tuesdays and Thursdays (send him an email to schedule an appointment.)

2.3 Web

- Web site for this course
- Dr. Love's GitHub name is THOMASELOVE.
- His Twitter handle is @ThomasELove

2.4 A More Complete Biography

Hi. 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 some awards for my teaching and my research.
- I have been teaching at CWRU since 1994, and have taught every type of CWRU student over the years, especially graduate 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 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 list of many 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 advisor was Paul Rosenbaum. I am married to a brilliant woman and we are raising two terrific sons, the elder of whom just finished his first semester of college. I live in Shaker Heights. In spare moments, I do community theater, and have appeared onstage with several local groups.

Chapter 3

Teaching Assistants

The teaching assistants for 431 this year are Bob Winkelman, Satyakam Mishra, Maher Kazimi, Zuxi (Terry) Cui and Xueyi (Julia) Zhang. 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 January 22.

3.1 Office Hours for TAs

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

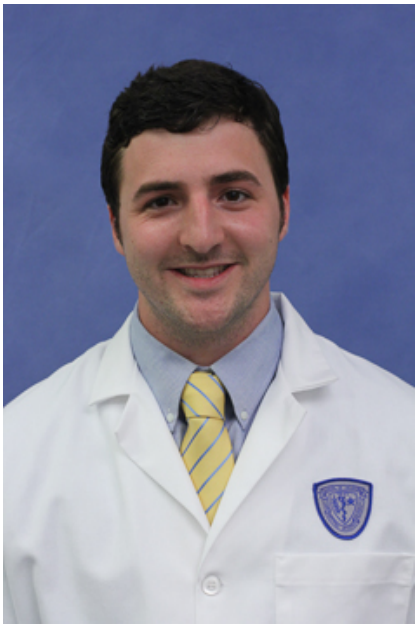
Teaching Assistant Office Hours are held in either WG-56 (Computing Lab) or WG-67 (Student Lounge) on the ground floor of the Wood building in the School of Medicine, so be sure to look in both places if you need help. The weekly schedule is:

Date	Time
Mondays	11:30 AM to 12:45 PM
Tuesdays	11:30 AM to 12:45 PM, 2:30 PM to 3:45 PM
Wednesdays	11:30 AM to 12:45 PM
Thursdays	11:30 AM to 12:45 PM, 2:30 PM to 3:45 PM
Fridays	11:30 AM to 12:45 PM

TA office hours will begin at 2:30 PM on Tuesday 2019-01-22. 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 information is also available in the Course Calendar.

3.2 Bob Winkelman



Bob Winkelman is a fourth year student in the M.D. program at the School of Medicine who is also enrolled in the M.S. program in Biostatistics. He received his undergraduate degree from Carnegie Mellon University where he majored in Chemical and Biomedical Engineering. Before coming to CWRU for medical school, Bob also worked for Epic, an electronic health record vendor, in Wisconsin. Bob 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 Cleveland Clinic Center for Spine Health. Bob enjoys playing flag football, exercising, cooking, reading, and listening to podcasts. Bob is doing an internship this semester with Dr. Love at Better Health Partnership helping to incorporate social determinant information based on geography. Bob is Co-Lead TA for 432.

3.3 Satyakam Mishra

Satyakam Mishra is in his second year in the M.S. program in Systems Biology and Bioinformatics. He completed his Bachelor's work in Biotechnology in India. He is primarily interested in the applications of statistical ideas in genetics research, and is currently analyzing conformational changes in protein structures

with varying ligand-receptor combinations. The 431 and 432 coursework has turned out to be immensely useful in his thesis work. Satyakam's hobbies include cooking, reading and watching sports, and he enjoys outdoor activities, such as hiking and trekking. Satyakam is Co-Lead TA for 432.

3.4 Maher Kazimi

Maher Kazimi is in his second year in the Masters of Public Health program's Population Health Research track. He is an international medical graduate from Jordan, and has practiced medicine in Jordan and Germany as a part of his internship in primary care medicine. He does research work in large data bases describing outcomes for transplant patients (with data from the United Network for Organ Sharing) and emergency department visits (from the Nationwide Emergency Department Sample.) He is also a part of the Health Data Matters team project at the Department of Population and Quantitative Health Sciences, and he is working with University Hospitals on his capstone project.

3.5 Zuxi (Terry) Cui

Zuxi (Terry) Cui is in his third year in the Ph.D. program in Epidemiology & Biostatistics. Before that, Terry was a research associate at the University of Arizona Cancer Center doing Epi-genetic works related to skin cancer. His interests cover but not limited to missing data, longitudinal modeling, Epi-genetics, and imputation. He served as TA for 432 one year ago mainly dealing with statistical and coding issues. Terry wants to be a young Biostatistician & Genetic Epidemiologist in a few years. He is also the father of a pup. and a kitten. and a fan of Asian classic and country music.

3.6 Xueyi (Julia) Zhang

Julia is in her first year in the Ph.D. program in Epidemiology and Biostatistics, in the Department of Population and Quantitative Health Sciences. She holds an MS in Statistics from CWRU's School of Arts

and Science. Julia's research interests include genetic epidemiology, statistical genetics, and machine learning.

Chapter 4

Deliverables and Grading

4.1 Timing and Deadlines

The Calendar is the most up-to-date resource for all deadlines in the course.

4.2 Grading

The final course grade is weighted as follows:

- 15% Class Participation / Group Work / Minute Papers
- 25% Six Homework Assignments
- 25% Two Quizzes
- 35% 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.

4.3 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.

4.4 Homeworks

There are 6 homework assignments this semester.

Details on those assignments are posted on the Course Homework Page.

4.5 Quizzes

Students are required to complete two quizzes, one in early March, and one in late April.

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

4.6 Projects

Students are required to complete two project assignments, one completed in mid-March (at the end of Spring Break), and the other at the end of the term.

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

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 . . .