DATA SCIENCE MODELING I

Description

Data Science Modeling I (STAT240) introduces students to reproducible data management, modeling, and analysis through a practical, hands-on case studies approach. Topics include the use of an integrated statistical computing environment, data wrangling, the R programming language, data graphics and visualization, random variables and concepts of probability, data modeling, and report generation using R Markdown with applications to a wide variety of data to address open-ended questions.

Learning Outcomes

STAT240 aims to provide a thorough introduction to the concepts and methods of statistical data science and data modeling. It is the first course in a two-course sequence. Students can expect to spend time each week viewing the posted lecture content, reading from the textbook or other assigned readings, working through examples, doing the homework exercises, and generally becoming skilled in statistical computing and thinking statistically.

Students who complete this course successfully will learn to:

- **wrangle data**: transform data, possibly from multiple sources, into a form convenient for analysis;
- **explore data**: visualize and summarize data, generate questions/hypotheses, and address them;
- ⓘ **ogram**: write R code using the RStudio integrated statistical computing environment to carry out data wrangling, graphical data exploration, and analysis that is reproducible;

- **model data**: provide low-dimensional summaries of data that capture signal and quantify the noise; assess the adequacy of the model; understand random variables and probability concepts associated with the models;
- **interpret data**: explain what can be inferred from the data analysis and make predictions;
- **communicate**: use R Markdown to integrate prose, visualizations, code, interpretation, and results;
- **collaborate**: work with other students to solve data challenges.

Furthermore, students will learn about:

- **statistical inference**: the construction and interpretation of confidence intervals and the calculation and interpretation of p-values for hypothesis tests for a number of settings including one- and two-sample proportions and means
- **simple linear regression**: the construction and interpretation of regression models for two quantitative variables

Course Materials

Textbook:
R for Data Science by Wickham and Grolemund
Statistics 240 Course Notes and Case Studies by Bret Larget

Software:
R: https://cran.r-project.org/
R Studio: https://www.rstudio.com/

Grading

The final course grade will be determined by a score made up from the *weighted* sources below.

Short Online Assessments (SOAs): Each week will include reading assignments and most will include an online quiz (SOA) which will examine your understanding of the assigned reading. Reading assignments will come from the online textbook and the course notes, and will be administered through Canvas. Lectures will assume that you have read and comprehended assigned reading. Anticipate spending approximately 30 - 60 minutes outside of class for each reading assignment and quiz.

Discussion Sessions: Weekly discussion sessions will include a discussion assignment. Discussion assignments are short group assignments meant to be completed during the 50-minute discussion period. Groups will be assigned early in the semester (prior to the third week of classes), and new groups will be assigned after the midterm exam. Only one group member

needs to turn in each group assignment, but all group members are expected to contribute to doing the work. Discussion assignments are intended to practice a single concept. Each discussion assignment will involve editing an R Markdown file to answer several questions, knitting the document to HTML, and uploading the knitted document to the Canvas web page.

Individual Assignments: In addition to group assignments in discussion sessions, there will be individual assignments, typically due on the Friday after they are assigned. These are assignments are longer than the group assignments, and each individual assignment may take 4–6 hours to complete.

Exams: There will be an midterm exam and a cumulative final exam.

Final group project: Each student will complete a project with an assigned group of students. The project will include the acquisition of a data from a novel source, an interesting question to address, the creation of an R Markdown document that contains a reproducible process of data manipulation, graphical exploration, modeling, analysis, and interpretation, using the appropriate methods from the course. The project will culminate in a report displayed as an HTML document.

Contact email: jjkehe@wisc.edu

Google Scholar GitHub ORCID Astrostatistics News

aut viam inveniam aut faciam