MATH8050: Data Analysis
Clemson University, Fall 2022

Please note that the syllabus may change if needed.

1 Course Schedule

Instructor: Pulong Ma, Assistant Professor, School of Mathematical and Statistical Sciences
Email: plma@clemson.edu
Course Time & Location:

- Section 001: Tuesday/Thursday: 8:00 – 9:15 AM EDT; Martin Hall M203
- Section 002: Tuesday/Thursday: 9:30 – 10:45 AM EDT; Martin Hall M203

Academic Calendar: https://www.clemson.edu/registrar/academic-calendars/

Lab Schedule (not required) *Lab with Tianqi Zhang, PhD Student, tianqiz@clemson.edu*: TBD

2 Course Learning Objectives

In this course, the following learning objectives will be met by the end of the semester:

1. Provide foundations on statistical learning from frequentist and Bayesian perspectives
2. Explore, visualize, and analyze data in a reproducible and shareable manner using statistical methods from the course
3. Gain experience in data wrangling and munging, building statistical models (linear regression, hierarchical regression, generalized linear regression), visualizing them, and interpreting them using programming languages such R or Python
4. Be able to check model assumption and perform model validation and comparison
5. Be able to apply and diagnose regression models (linear regression, hierarchical regression, generalized linear regression) using both frequentist and Bayesian approaches
6. Be able to select and apply appropriate variable selection techniques using frequentist and Bayesian approaches
7. Be able to derive any mathematical results (e.g., proving theorems or deriving formulas) and interpret them
8. Work on problems and case studies inspired by and based on real-world questions and data
9. Learn to effectively communicate results through written assignments, exams, and projects using R Markdown and LaTeX

3 Course Community

Clemson Community and Ethical Standards While you are a student here at Clemson University, it is our expectation that you uphold the standards and core values of the Clemson University Community. We have defined our core values as Integrity, Honesty, and Respect. As a new student, whether you are an incoming freshman, transferring from another institution, or a new graduate student, Clemson University expects you to adopt these core values as your own. Also, the standards of our community can be found in the Student Code of Conduct.

Academic Integrity As members of the Clemson University community, we have inherited Thomas Green Clemson's vision of this institution as a "high seminary of learning." Fundamental to this vision is a mutual commitment to truthfulness, honor, and responsibility, without which we cannot earn the trust and respect of others. Furthermore, we recognize that academic dishonesty detracts from the value of a Clemson degree. Therefore, we shall not tolerate lying, cheating, or stealing in any form.
Please abide by the following as you work on assignments in this course:

1. You may not discuss or otherwise work with others on the exams. Unauthorized collaboration or using unauthorized materials will be considered a violation for all students involved.

2. Reusing code: Unless explicitly stated otherwise, you may make use of online resources (e.g. StackOverflow) for coding examples on assignments. If you directly use code from an outside source (or use it as inspiration), you must explicitly cite where you obtained the code. Any recycled code that is discovered and is not explicitly cited will be treated as plagiarism.

3. On individual assignments, you may not directly share code or write up with other students. On team assignments, you may not directly share code or write up with another team. Unauthorized sharing of the code or write up will be considered a violation for all students involved.

Any violations in academic honesty standards as outlined in the Clemson Community and Ethical Standards and those specific to this course will automatically result in a 0 for the assignment and may result in a failing grade for the course. Violations will be reported to the Office of Community and Ethical Standards for further action.

Accessibility Clemson University values the diversity of our student body as a strength and a critical component of our dynamic community. Students with disabilities or temporary injuries/conditions may require accommodations due to barriers in the structure of facilities, course design, technology used for curricular purposes, or other campus resources. Students who experience a barrier to full access to this class should let the professor know, and make an appointment to meet with a staff member in Student Accessibility Services as soon as possible. You can make an appointment by calling 864-656-6848, by emailing `studentaccess@lists.clemson.edu`. Students who receive Academic Access Letters are strongly encouraged to request, obtain, and present these to their professors as early in the semester as possible so that accommodations can be made in a timely manner. It is the student's responsibility to follow this process each semester. You can access further information at the campus services website.

CU Title IX The Clemson University Title IX statement: Clemson University is committed to a policy of equal opportunity for all persons and does not discriminate on the basis of race, color, religion, sex, sexual orientation, gender, pregnancy, national origin, age, disability, veteran's status, genetic information or protected activity in employment, educational programs and activities, admissions and financial aid. This includes a prohibition against sexual harassment and sexual violence as mandated by Title IX of the Education Amendments

of 1972. This Title IX policy is located on the Campus Life website. Ms. Alesia Smith is the Clemson University Title IX Coordinator, and the Executive Director of Equity Compliance. Her office is located at 223 Brackett Hall, 864.656.0620. Remember, email is not a fully secured method of communication and should not be used to discuss Title IX issues.

4 Course Activities

Communication I will email announcements through Canvas to ensure everyone has current information. Please check your email regularly. I will send out surveys periodically to gauge how the class is going, so please respond to these as I greatly appreciate your feedback to improve the course!

Teaching Assistant A Graduate Teaching Assistant is available to help and assist you throughout the course!

1. Tianqi Zhang, PhD Student in School of Mathematical and Statistical Sciences, tianqiz@clemson.edu

Office Hours We will make adjustments to OH based upon student needs/feedback.

1. Professor Ma, Tuesday/Thursday: 10:45 – 11:45 AM EDT
2. Tianqi Zhang, PhD Student, tianqiz@clemson.edu, TBD

Lecture Component Lectures will be conducted as in-person class. The majority of class will be dedicated to the following:

1. Prof. Ma will go through the concepts and mathematical derivation in class, highlighting the most important parts, providing clarifications, solutions/advice, and more advanced insights.
2. Prof. Ma will take questions or those that are emailed to him by Monday at 10 AM and Wed at 10 AM.
3. Prof. Ma will provide exercises for the class to work through and an interactive environment.

Lab Component The lab sessions will be either in classroom or on Zoom during the scheduled lab period.

1. Students are encouraged to post lab questions in advance to Canvas regarding lab, which should be posted no later than each Thursday at 5 PM.
2. The TA will go over all lab tasks, providing clarifications, solutions/advice, and more advanced insights during next week.
3. The TA will take questions or go over those posted to Canvas.
4. Attendance is highly encouraged in order to learn tips and tricks and in order to get hints/advice which are important for homework and exams.

Getting Help If you have a question during lecture or lab, feel free to ask it! There are likely other students with the same question, so by asking you will create a learning opportunity for everyone.

The Teaching Assistant is here to help you be successful in the course. You are encouraged to attend office hours to ask questions about the course content and assignments. A lot of questions are most effectively answered *in person*, so office hours are a valuable resource. Please use them!

Outside of class and office hours, any general questions about course content or assignments should be posted on Canvas.

If you send an email to myself, your email subject should start with **MATH8050 Section 001: ...** or **MATH8050 Section 002: ...**. If you believe your question can be addressed by the TA, please make sure to CC your email to the TA for fastest response.

No help from the instructor or TA will be offered for the homework that is due on the due date. It is strongly recommend that you should work on the homework in advance. Given a large volume of emails received by the instructor or TA, they need to respond emails on a "first come, first served" basis. please ask questions early.

5 Prior Knowledge, Course Expectations, and Grading Policies

Prior Knowledge This course assumes that students have backgrounds in undergraduate level linear regression and elementary probability. In particular, students are expected to have a solid background in Statistics for Science & Engineering (MATH 3020) and Theory of Probability (MATH 4000/6000), or Statistics for Science & Engineering II (MATH 4020/6020) and Probability (MATH 8000). These will be building blocks for course topics, and very little review will be provided in this course. If you are unsure what prior knowledge is expected, please refer to the following resources, and review any gaps in knowledge before the start of the semester. Below is a list of available recourses

1. Statistics for Engineers and Scientists: video available at YouTube
2. Linear Regression and Modeling: available at Coursera
3. Bayesian Statistics: available at Coursera
4. Linear Algebra: `http://www.stat.columbia.edu/~fwood/Teaching/w4315/Fall2009/lecture_12`
5. R programming. Students are expected to have a solid foundation of R programming prior to the first day of lecture. For a review of R programming, please see Lecture 0 and recourses on Canvas.
6. R Markdown. Students are expected to know how to use R Markdown to generate assignments and reports.
7. LaTeX. Students are expected to know how to use LaTeX language in R markdown for completing assignments and writing final project report.

Remark: This course will be very difficult without a strong foundation in the topics above, so please do review these in advance.

Expectations

1. Students are expected to be very familiar with R and are expected to know how to use R markdown.
2. All homework, reports, and take home exams (if applicable) should be submitted in Markdown **.Rmd** and **.pdf** format.[1]
3. Please name your reports using the naming convention in the following example. As an example, please using the following naming convention **HW1-FirstName-LastName.Rmd**. Please also refer to the file **HW-template.Rmd** for detailed instructions.
4. All homework submissions must be made through **Canvas**.
5. **When submitting to Canvas, only one file must be uploaded.** Please zip together all materials for your homework assignment and upload the zipped file. As an example, please upload **HW1-FirstName-LastName.zip**, which should be a folder that contains **HW1-FirstName-LastName.Rmd** and **HW1-FirstName-LastName.pdf**.[2]
6. Your homework and reports are expected to be reproducible and compile for full credit. **If your R markdown file cannot reproduce the submitted pdf file (including the situation where it fails to run), the grade of that homework assignment will receive 10% penalty.**
7. Students are expected to keep up with the reading in the course and have read before they come to class. Finally, if students find typos on the slides, please write them down with the slide and typo and give them to Professor Ma for a timely correction.
8. **Attendance is expected of all students given that homework and exam material will come from both class and lab.**
9. There will be between 12 homework assignments during the course of the semester. One lowest homework grade will be dropped.

All homework involving analysis and code must be submitted to Canvas using R/Rmarkdown. Specifically, your homework must be reproducible. Your homework must be included as one file, therefore, please zip your files and submit all the files using a .zip extension. If you are unsure of how to do this, please see your TA or instructor during the first week of class during OH. Submissions via email to the TA or instructor will not be accepted for credit.

If you have not downloaded LaTeX, you will need this in order for your .Rmd file to compile to a .pdf file. To install LaTeX, please see https://www.latex-project.org/get/. This will direct you to options depending on if you are a Windows, Mac, or Linux user.[3] At least one student in the class has LaTeX working on Window using https://tug.org/texlive/acquire-netinstall.html or https://miktex.org/download. If you want to do a full install of LaTeX for Windows, see https://tug.org/texlive/doc/texlive-en/texlive-en.html#installation.

[1] In case you use Jupiter notebook/Lab, you will need to submit **.ipynb** and **.pdf** files.

[2] If you are working with data in a homework assignment, please make sure to also attach the data and also make sure that when you call the data in your markdown file, there are no hard coded commands. For example, make sure you do not set your working directory because we won't be able to reproduce your file.

[3] For Mac users, I like this option here: http://www.tug.org/mactex/. Make sure to install the full version and not the small version. For a compiler, I like TexShop.

Prerequisites You are expected to have all pre-reqs to be in the course (see Prior Knowledge section above). Students are expected to be very familiar with `R` and are **encouraged** to have learned `LaTex` by the end of the course.

- Required textbooks: Please make sure you have a copy of the book as there will be required reading throughout the course.
  - *An Introduction to Statistical Learning with Applications in R, Second Edition*, James, Gareth, Witten, Daniela, Hastie, Trevor, and Tibshirani, Robert, 2021. New York: Springer. The pdf version of this book is freely available online at Springer or at `https://www.statlearning.com/`. I will refer to this book as "ISL" throughout the course.
  - *Bayesian Data Analysis, Third Edition*. Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., & Rubin, D.B. (2013). CRC press. The pdf version of this book is freely available at `http://www.stat.columbia.edu/~gelman/book/`. I will refer to this book as "BDA" throughout the course.
  - *A First Course in Bayesian Statistical Methods*, Peter D. Hoff, 2009, New York: Springer. The pdf of this book can be purchased at Springer. I will refer to this book as "PH" throughout the course.
- Optional supplementary text:
  - *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*, Hastie, Trevor, Tibshirani, Robert, and Friedman, Jerome, 2009. Springer-Verlag. The pdf of the book is freely available at the book website. I will refer to this book as "ESL".
  - *Statistical Inference, Second Edition*. Casella and Berger `https://mybiostats.files.wordpress.com/2015/03/casella-berger.pdf`
  - *The R Cookbook*, `http://www.cookbook-r.com/`.
  - *R for Data Science*, `https://r4ds.had.co.nz/index.html`.
  - *R Markdown: The Definitive Guide*, `https://bookdown.org/yihui/rmarkdown/`.
  - *10-minute tutorial on R Markdown*, `https://www.markdowntutorial.com`.

Grading Policy: The following grading policy will be used for this class.

Table 1: Grading Policy:

Homework	30%
Exam I: Thursday, Oct 6, 2022	20%
Exam II: Thursday, Nov 17, 2022	20%
Final Project	30%

Homeworks will be given on a weekly basis with **due dates at 12PM noon every Wednesday**. They will be based on both lecture and lab. **To accommodate unexpected events, one lowest homework grade will be dropped at the end of the semester**.

There will be two exams during the semester and one final project. The purpose of the exams are an opportunity to assess the knowledge and skills you've learned. They will include both the conceptual and mathematical

aspects of data analysis. Exams will be timed and must be completed during the time period specified.

An overall score of s will result in a grade of:

A if $90 \leq s \leq 100$
B if $80 \leq s < 90$
C if $70 \leq s < 80$
D if $60 \leq s < 70$
F if $0 \leq s < 60$

Exam dates cannot be changed and no make-up exams will be given. If extenuating circumstances prohibit you from taking an exam, please let Professor Ma know before the start of the exam. Please remember that you must complete the final project to pass the course.

Students are not to discuss any contents of any examination until after the exam grades are released back to them either in class or via Canvas. More specifically, students should not speak to anyone except the course instructor until exam grades are released to the entire class. This includes but is not limited to talking to other students, text, chat forums, and other means of communications where exam information could be shared to another student. Any student that does not follow this policy will be in violation of the Student Conduct Code.

Late Work **No late assignments will be accepted after the due date under any circumstances**. To accommodate unusual circumstances that prevent the completion of homework, one lowest homework grade will be dropped from the evaluation of the homework average.

6 Exams and Final Project

Exams The two exams will focus both on programming and mathematical skills. Each exam will consist of two parts: In Part One, the exam will test mathematical skills in deriving key statistical/mathematical formulas and results during the class time. A calculator permitted. This part accounts for 60% of the overall exam score and is closed book; In Part Two, you will be asked to use programming language such as R to complete all the problems. A report in pdf format (along with its .Rmd file or .ipynb) must be submitted by the end of the exam day. This part accounts for 40% of the overall exam score.

Final Project In the finial project, it is highly encouraged for a student to team up with up to three other students. Each team should submit a preliminary report, a mid-stage report, and final project report for the project. Along with each report, you should also submit a zipped file including a R Markdown file with code and data to ensure that all your figures can be reproduced. If your R Markdown file fails to run or cannot reproduce your results in the report, your score will get 10% penalty. At the end of the semester, each team will have 15 minutes to present their work. The template to write your report is available on Canvas.
The preliminary report is due on Oct 10. In the preliminary report,

- the team should choose a publicly available dataset and describe the scientific problem that the team is trying to solve;
- the team should perform exploratory data analysis (EDA) and describes initial findings from EDA;
- the team should make arguments to support the statistical models/techniques that will be selected to analyze the dataset.

- the report should be 1-2 pages long excluding any graphs, code, and references.

The mid-stage report is due on Nov 18. In the mid-stage report,

- the team should describe in detail the statistical models/techniques used and why they are appropriate to use.
- the team should implement at least two statistical models/techniques, perform model diagnostics, and compare findings with different methods.
- the report should be 3-4 pages long excluding any graphs, code, and references.

The final project report is due on Dec 8. In the final project report,

- both frequentist and Bayesian approaches should be used to analyze the real dataset.
- all statistical models (along with the assumptions) should be fully described and why they are appropriate for the dataset should be fully addressed.
- limitations of the adopted methods/models for the real dataset should be discussed.
- the report should be 5-6 pages long excluding any graphs, code, and references.

The evaluation of final project will be based on readability of project reports (10%), the preliminary project report (10%), the mid-stage project report (20%), the final project report (30%), and the final project presentation (30%).

7 Important Dates

1. Aug 24: Classes Begin
2. Sep 6: Drop/Add Ends
3. Oct 28: Last Day to withdraw without final grades
4. Nov 7 - Nov 8: Fall break
5. Nov 23 - Nov 25: Thanksgiving holidays
6. Dec 6, 8: Final project presentation

Exam #1	Oct 6, 2022
Exam #2	Nov 17, 2022
Preliminary Project Report	Oct 10, 2022
Mid-Stage Project Report	Nov 18, 2022
Final Project Presentation	Dec 6, 8, 2022
Final Project Report Deadline	Dec 12, 2022

Tentative Schedule

Week	Topic	Assignment
1	Introduction	Reading: David Donoho (2017) "50 Years of Data Science." ISL Ch1-2 & PH Ch1-2.
2	Data wrangling with Tidyverse Lab 1	Reading: See Canvas.
3	Data wrangling with Tidyverse Lab 2	Reading: See Canvas. **Homework 1 due on 9/7/2022.**
4	Linear regression: Least squares, MLE Lab 3	Reading: ISL Ch3.1-3.2 & ESL Ch3.1-3.2. **Homework 2 due on 9/14/2022**.
5	Residual diagnostics & Logistic regression Lab 4	Reading: ISL Ch3.3. **Homework 3 due on 9/21/2022**.
6	Monte Carlo simulation Lab 5	Reading: PH Ch4, 6 & BDA Ch10, 11. **Homework 4 due on 9/28/2022**.
7	Metropolis algorithm and Gibbs sampling Lab 6	Reading: PH Ch5, 7 & BDA Ch2.5 **Homework 5 due on 10/5/2022**. **Exam 1 on 10/6/2022.**
8	Bayesian linear regression: g-prior and model selection Lab 7	Reading: PH Ch9 & BDA Ch2.8-3.6 **Homework 6 due on 10/12/2022**.
9	Bayesian model criticism and model selection Lab 8	Reading: BDA Ch6, 7. **Homework 7 due on 10/19/2022.**
10	Cross-validation & bootstrap Lab 9	Reading: ISL Ch5.1. **Homework 8 due on 10/26/2022**.
11	Model selection: Subset selection Lab 10	Reading: ISL Ch6.1 & ESL Ch3.2. **Homework 9 due on 11/2/2022**.
12	Regularization: Ridge regression and the lasso Lab 11	Reading: ISL Ch6.2, 6.4 & ESL Ch3.4. **Homework 10 due on 11/9/2022**. **Fall break: Nov 7-8, 2022**.
13	Bayesian shrinkage: Spike and slab priors Lab 12	Reading: TBD. **Homework 11 due on 11/16/2022**. **Exam 2 on 11/17/2022**.
14-15	Gaussian process and tree-based methods	Reading: TBD. **Homework 12 due on 11/30/2022**. **Thanksgiving holidays: Nov 24-25, 2022**.
16	Final Project Presentation	