



STA 4241

Spring 2021

Statistical Learning in R

**Instructor:** Joseph Antonelli

E-mail: [jantonelli@ufl.edu](mailto:jantonelli@ufl.edu)

Office hours: Time TBD. Location via Zoom for students in either the online or in-person sections.

**Teaching Assistant:** Yisen Jin

E-mail: [y.jin@ufl.edu](mailto:y.jin@ufl.edu)

Office hours: Time TBD. Location via Zoom for students in either the online or in-person sections.

**Course Website:** [e-Learning](#)

**Course Prerequisites:** STA 4322 & STA 4210 & MAS 4115 (Or linear algebra equivalent)

**Course lectures:** Attendance for lectures is strongly recommended and they will be given MWF from 10:40am to 11:30am. For those taking the class in-person, this will take place in TUR 1105, while the online students will attend via Zoom. We will likely be recording the lectures and posting them to the course website, but this is our first semester working with HyFlex teaching, so please attend the live session at least until that is figured out properly.

**Course Notes/Material:** Notes for the week will be posted at the beginning of each week on the course website. These should contain nearly all of the material that we cover in class, however, I will present some additional material in the lectures that is not posted on the course website.

**Software:** We will be using the R software language throughout. R is free and should be easy to download on your personal computer. I also highly recommend running R through RStudio, though it is not a requirement. If you have any problems downloading R or RStudio, feel free to talk to myself or the TA. If you do not have access to a computer, please reach out to me via email.

**Required Text:** James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) An Introduction to Statistical learning with Applications in R. Springer.

**ISBN-13:** 978-1461471370

**Course Description:** Overview of the field of statistical learning. Topics include linear regression, classification, resampling methods, shrinkage approaches, nonlinear regression, tree-based methods, support vector machines, and neural networks. We will

cover many aspects of these approaches including conceptual, theoretical, and applied aspects. Approaches will be illustrated and implemented in R.

**Course Objectives:** The goal of this course is to teach the theoretical underpinnings of a number of advanced and commonly used statistical learning techniques. We will review classical statistical techniques such as linear and logistic regression before covering more advanced statistical techniques such as classification, regularization, nonlinear regression, and other machine learning approaches. The implementation of all approaches in the R statistical software will be taught throughout. By the end of the course, students should be familiar with a wide range of statistical methodologies that are widely used in practice, and should be able to apply these approaches to data sets.

### Homework

There will be a homework assignment every two weeks and it will be due via Canvas submission at the end of that week.

### Exams

You will have one take-home exam that is to be assigned roughly in the middle of the semester.

**Project:** Students will be expected to complete a written project at the end of the semester and present their findings to the class. The grade for the final project will consist of three main components: statistical modeling decisions and code, a written report, and an oral presentation. The first component is based on whether the student used appropriate statistical tools for the setting and objectives, whether the coding done was efficiently and correctly, and whether the conclusions are consistent with their analysis. The written report will state the objectives of the study, describe data collection, describe the statistical model used, explain any assumptions required by the analysis, and provide conclusions for the main study questions. The oral presentation will be a ten minute presentation over Zoom that should cover the key components of the written report. Presentations should clearly state the objectives of the project, while using visualizations to illustrate the main results and findings of the project.

### Grade Distribution

Homework	20%
Midterm	40%
Project	40%

**Letter Grade Assignment:** Grades will be assigned as follows: 90-100, A; 87-89.9, A-; 84-86.9, B+; 80-83.9, B; 77-79.9, B-; 74-76.9, C+; 70-73.9, C; 67-69.9, C-; 64-66.9, D+; 60-63.9, D; 55-59.9, D-; 0-55, F

The numeric scores will be rounded to the nearest tenth.

**Make up Policy:** Requirements for class attendance and make-up exams, assignments, and other work in this course as well as policies regarding absences, religious holidays, illness and student athletes are consistent with [UF Attendance Policies](https://catalog.ufl.edu/UGRD/academic-regulations/attendance-policies/) (<https://catalog.ufl.edu/UGRD/academic-regulations/attendance-policies/>)

### **Dropping and Withdraw**

For late course drops and course withdrawals please visit <https://catalog.ufl.edu/UGRD/academic-regulations/dropping-courses-withdrawals/>

### **Incomplete**

An incomplete grade may be assigned at the discretion of the instructor as an interim grade for a course in which the student has completed a major portion of the course with a passing grade, been unable to complete course requirements before the end of the term because of extenuating circumstances, and obtained agreement from the instructor and arranged for resolution of the incomplete grade in the next term. Instructors are not required to assign incomplete grades. For complete details please visit [CLAS incomplete grade policies and forms](https://www.advising.ufl.edu/academicinfo/clas-policiesprocedures/incomplete-grades/). (<https://www.advising.ufl.edu/academicinfo/clas-policiesprocedures/incomplete-grades/>)

### **Accommodating Students with Disabilities**

Students requesting accommodation for disabilities must first register with the Dean of Students Office. The Dean of Students will provide documentation to the students who must then provide this documentation to the instructor when requesting information. You must submit this documentation prior to submitting any assignments for which you are requesting accommodation.

**Academic Misconduct:** Students are held accountable to the [UF Honor Code](https://sccr.dso.ufl.edu/process/student-conduct-code/). (<https://sccr.dso.ufl.edu/process/student-conduct-code/>)

**Evaluations:** Students are expected to provide feedback on the quality of instruction in this course by completing online evaluations at <https://evaluations.ufl.edu>. Evaluations are typically open during the last two or three weeks of the semester, but students will be given specific times when they are open. Summary results of these assessments are available to students at <https://evaluations.ufl.edu/results/>.

**Additional resources:** Any additional resources including academic support or information technology can be found at <https://www.ufl.edu/about/offices-services/>

**Privacy statement regarding online lectures:** Our class sessions may be audio-visually recorded for students in the class to refer back and for enrolled students who are unable to attend live. Students who participate with their camera engaged or utilize a profile image are agreeing to have their video or image recorded. If you are unwilling to consent to have your profile or video image recorded, be sure to keep your camera off and do not use a profile image. Likewise, students who un-mute during class and participate verbally are agreeing to have their voices recorded.

If you are not willing to consent to have your voice recorded during class, you will need to keep your mute button activated and communicate exclusively using the "chat" feature, which allows students to type questions and comments live. The chat will not be recorded or shared.

As in all courses, unauthorized recording and unauthorized sharing of recorded materials is prohibited.

### **Weekly breakdown of the material:**

#### Week 1

- Introduction to Statistical Learning

#### Week 2

- Review of linear regression

#### Week 3

- Logistic regression
- Linear discriminant analysis

#### Week 4

- Quadratic discriminant analysis
- Maximal margin classifier

#### Week 5

- Support vector classifier
- Support vector machines

#### Week 6

- Cross validation
- The bootstrap

#### Week 7

- Linear model selection and regularization
- Subset selection

#### Week 8:

- Take home exam

#### Week 9:

- Shrinkage methods
- Ridge regression
- The lasso

#### Week 10

- Dimension reduction methods
- Principle components analysis, Principal components regression
- Partial least squares

#### Week 11

- Nonlinear models
- Polynomial regression - regression splines
- Smoothing splines

#### Week 12

- Tree-based methods
- Decision trees

#### Week 13

- Bagging
- Random forests
- Boosting

#### Week 14

- Ensemble approaches
- Neural networks

#### Week 15

- Project presentations