Syllabus: Statistics 8330, Data Analysis III

Fall 2020

About the Instructor

Name:	Chris Wikle
Phone number:	573-882-9659
Email address:	wiklec@missouri.edu
Campus office (if applicable):	146 Middlebush Hall
Virtual Office Hours:	from end of class to 3:45 M,W 2:00-3:00pm, F
In-Person Office Hours:	none

As your instructor, I will:

- *work hard to convey the course goals and learning objectives in a respectful and positive environment*
- *make myself available to answer questions, provide guidance, and offer support.*
- *respond to questions within 2 business days (usually faster).*
- *monitor and facilitate discussions, Monday–Friday.*
- *challenge you, motivate you, and help foster a learning community*

Welcome

Welcome to the third semester in our graduate data analysis sequence! By now, you have been exposed to many of the most important general statistical methods. This class will take you a couple of steps beyond, both in terms of the philosophy of statistical modeling and exposure to many new methods that have become popular in statistical learning in the last two decades. Specifically, the course will focus on advanced data analysis and methodology associated with statistical learning (data mining, inference and prediction). When you finish this course, you will be able to step into almost any real-world data analysis situation and have multiple approaches in your toolkit by which to solve those challenges.

This is my favorite course to teach because it provides connections to all of the methods we have learned before. But, as a broad survey course, this class covers a large amount of material and we move through the methods fairly quickly. You will be expected to know the material in the class notes and to have read the recommended reading material. More importantly, you will be responsible for a fairly large amount of statistical programming using "R", including going through R labs that are provided in the class textbook on your own.

Homework is an important part of the course. It will largely include exercises associated with the computer labs. There will be approximately 8 homework assignments. In addition, there will be 3 projects related to real-world data. At least two of these will be a competitive project (details forthcoming).

Instructor Bio

I am a Curators' Distinguished Professor and Chair of Statistics. I received a PhD co-major in Statistics and Atmospheric Science in 1996 from Iowa State University. I was a research fellow at the National Center for Atmospheric Research from 1996-1998, after which I joined the Statistics Department here at Mizzou. My research interests are in spatio-temporal statistics applied to environmental, ecological, geophysical, agricultural and federal survey applications, with particular interest in dynamics. My work has been concerned with formulating computationally efficient deep hierarchical Bayesian models motivated by scientific principles, with more recent work at the interface of deep neural models in machine learning. I enjoy solving real-world problems with cool statistical methods!

Teaching Philosophy

My lecture style is more interactive than most Statistics classes you have had in the past, although this is the first time I've taught this course online. My goal is to help you understand the key concepts and that is the focus of my lecture. I also believe firmly that ***you learn data analysis by doing data analysis****, and this experience (through homework and projects) will be the most important and valuable part of the class. It will also require a large time commitment from both of us. But, if you put in the work, I will do my best to help you get through the challenging parts of the course and you will come out the other side a much better data analyst than when you started!*

Course Materials

Required:

An Introduction to Statistical Learning with Applications in R, 2013 (2017), James, Witten, Hastie, Tibshirani; Available online for free.

The Elements of Statistical Learning, 2nd Edition, 2009, Hastie, Tibshirani, Friedman; Available online for free.

Class lecture notes (provided)

Required Software/Technology:

We will use Canvas and Zoom in this class. For computing, we will use the "R" programming language exclusively in this class. A wealth of information on R can be found at: http://www.r-project.org/

I recommend the manual "An Introduction to R" as a place to start. You can also download the software to your own machine from this same site.

Revised August 21, 2020

Digital Literacy and Technical Support

Minimum Technology Requirements:

At a minimum, you will need the following software/hardware to participate in this course:

- Stable DSL or Cable Internet connection or a connection speed no less than 6 Mbps.
- Computer with an updated operating system (e.g. Windows, Mac, Linux)
- Modern web browsers (Apple Safari, Internet Explorer, Google Chrome, Mozilla Firefox)
- Minimum Processor Speed of 1 GHz or higher recommended
- Media player such as VLC Media Player
- Adobe Flash player (free)
- Adobe Reader or alternative PDF reader (free)
- R software
- A webcam and/or microphone is highly recommended.

Note: This list represents the basic, minimum technology requirements for all learners in the course. Other technologies, such as online collaboration tools, may be required depending on decisions made within your peer group and decisions you make later in the course.

Technical Support

Problems with your computer or other technology issues are not an excuse for delays in meeting expectations and missed deadlines for the course. If you have a problem, get help in solving it immediately.

Mizzou Tech Support can provide step by step instructions on using Canvas tools, help you log in, or troubleshoot issues.

- Email techsupport@missouri.edu
- Call (573) 882-5000
- Visit http://doit.missouri.edu

If you are having difficulty with a technology tool in Canvas, consider visiting the Canvas Student Guides, which has overviews of each tool and tutorials on how to use them.

Minimum Technical and Digital Information Literacy Skills:

To excel in this course, you should have these incoming skills:

- The ability to download, edit, and save Word documents.
- The ability to effectively use peripheral computer components, including speakers, webcams, and microphones.
- The ability to download and install software on your personal computer and/or install apps on your mobile devices.
- The ability to perform a variety of functions within Canvas, including but not limited to:
- uploading and submitting Word documents;
- posting text-based comments on a discussion forum thread and replying to peers' posts;
- posting Announcements and/or creating and updating Pages within the Groups area in Canvas;

- opening files within the LMS for on-screen viewing and/or downloading them for offline viewing;
- accessing the Grades area;
- accessing the Calendar tool;
- updating your Account information (e.g. profile/bio, notification preferences, contact information); and
- using the Inbox feature to communicate with your instructor by email within the LMS.
- The ability to use online libraries and databases to locate and gather appropriate information.
- The ability to properly cite information sources.
- The ability to prepare a presentation of research findings.

If this is your first online course, complete the MU Canvas Online Student Orientation available in your Canvas course list. The orientation course is always available for reference.

About this Course

This course is 100% online, which means that all course-related activities will take place, in some manner, online. The majority of our interactions will be at pre-specified (we call this "synchronous"). Everything you need will be made available through the course and accessible via the Modules area. You will be expected to participate in the Zoom lectures. A couple of rules about the Zoom lectures: when you are talking, you will have your video on (if you cannot for some reason, please discuss with me); when we have general discussion, everyone's video will be on. During my lecture, only my video will be on.

The course will include regular lectures (30-60 min typically) covering the provided class notes. You will be expected to go through the R programming labs offline. I will hold virtual office hours from the end of lecture (which is usually before the end of class) to 3:45 on Monday and Wednesday, and then from 2-3 on Friday.   There will be three projects through the class (including at least one Kaggle competition).

Time Requirements

This is an active online course that requires 3 hours of your time each week for lecture and R labs, plus the time it takes you to read or watch the required materials, complete the assignments, and attend office hours. That means that you need to plan to spend a minimum of 15 hours every week on activities related to this course. It is essential that you access the course site regularly to read announcements and check for new assignments and support material. While it is vitally important to stay connected, it is also important to take a break. Keep this in mind when mapping out your semester schedule.

Course Description:

Statistics 8330: An introduction to data analysis techniques associated with supervised and unsupervised statistical learning. Resampling methods, model selection, regularization, generalized additive models, trees, support vector machines, clustering, nonlinear dimension reduction.
Prerequisites: STAT 8320

The course learning objectives will be conveyed through synchronous lectures, homework exercises, one online exam, and three data analysis projects (at least one of which will be a competition).

Goals of the Course:

This course will extend the methods you learned in STAT 8310 and STAT 8320 to more modern statistical learning methodologies. It will also solidify the concepts you learned in those courses by providing a deeper philosophy of data analysis. Specific course goals are:

- Understand the importance of tradeoffs between variance and bias in statistical modeling and how this is manifested in model flexibility and interpretability
- Understand and implement regularization in statistical models
- Understand the importance of cross-validation in assessing model accuracy
- Become proficient in the motivation and use of flexible regression models based on basis functions, kernels, local regressions, generalized additive models, and Gaussian processes
- Become proficient in the motivation and use of tree-based regression and classification models: CART, MARS, Bagging, Boosting, Random Forests
- Receive an introduction to the motivation and practice of using basic neural networks and deep learning
- Understand the motivation and application of support vector machines
- Understand extensions of basic methods in the presence of high-volume data
- Understand and use unsupervised statistical methods: cluster analysis, PCA, and nonlinear dimension reduction
- Understand and use hidden Markov models for complex time dependent processes
- Receive an introduction to latent process models, particularly latent Dirichlet allocation models for text analysis
- Applications to real-world problems through competitions and novel applications

How to Succeed in this Course

- *Attend the lecture*
- *Come to office hours and ask questions*
- *Read the material to be covered each day: re-read the previous lecture (and be prepared for questions); read the lecture notes for the day's lecture; read the relevant book material for that day's lecture*
- *Stay connected. Plan to log into Canvas at least once a day Monday–Friday to check announcements, check your Inbox (or, email), and keep up with the discussions (if there are any).*
- *Participate early and often in the discussions – particularly the real-time ones!*
- *Submit assignments on time.*
- *Participate in the online office hours.*
- *Read and use (or, at least, consider) the feedback you receive.*
- *Ask questions. Ask questions of me, your course peers, and your department colleagues.*
- *Allow yourself to be stretched, but not stressed. Use the support resources available to you when you need them.*

Assessment/Grading

The course grade will be determined from:

- *Homework assessments: 40%*
- *Midterm exam: 15%*
- *Data Analysis Projects: 40%*
- *Class participation (discussion, office hour help sessions): 5%*

The course will make use of a plus/minus grading system and will the grading scale will be "curved" (note: no matter the curve, if your class percentage is at or above 90% you will receive an A or A-).

Feedback and Grading Timeline:

I will strive to return graded homework and the mid-term exam within one week of assignment. Projects may take a little longer. You can review your grade and any feedback by selecting the Grades tab in the course menu.

Assignment Descriptions

- *Homework in this class are the key to learning the material as you have to do data analysis to learn data analysis. After going through the R labs in the book, you will work a few problems from the book, and then I will assign a different example relevant to this material. The assignment expectations are specific to each homework, but I will expect your writeup to be in R Markdown (or a similar package that integrates code and text). You can find an Introduction to R Markdown at this link. You should expect feedback and graded homework returned approximately one week after the due date.*
- *Projects will provide you an opportunity to analyze a "real-world" dataset with a specific goal. At least one of the three projects will be assigned from a Kaggle competition. You should expect feedback and graded projects within 1-2 weeks of the due date.*

Course Policies

For University policies and support resources, please click the Support & Policies link in the course menu.

Online Class Netiquette

Your instructor and fellow students wish to foster a safe online learning environment. All opinions and experiences, no matter how different or controversial they may be perceived, must be respected in the tolerant spirit of academic discourse. You are encouraged to comment, question, or critique an idea but you are not to attack an individual.

Revised August 21, 2020

Our differences, some of which are outlined in the University's nondiscrimination statement, will add richness to this learning experience. Please consider that sarcasm and humor can be misconstrued in online interactions and generate unintended disruptions. Working as a community of learners, we can build a polite and respectful course ambience.

Academic Integrity/Plagiarism
The Support & Policies link in the course menu includes a full, current version of the University's Academic Integrity policy. In this class, you are allowed to collaborate on homework, but I expect you to submit your own (independent) solutions, code, and write-up. You will not be allowed to collaborate on the midterm exam. Project collaboration will be discussed specifically for each project.

Intellectual Property
The Support & Policies link in the course menu includes a full, current version of the University's Intellectual Property policy. I expect that you will not share any of the course notes, homework questions/solutions, or recordings from this class without my permission.

Title IX Policies
Guidelines: The Support & Policies link in the course menu includes a full, current version of the University's Title IX policy

COVID-19 Statement

MU cares about the health and safety of its students, faculty, and staff. To provide safe, high-quality education amid COVID-19, we will follow several specific campus policies in accordance with the advice of the Center for Disease Control and Boone County health authorities. This statement will be updated as information changes.

- **If you are experiencing any COVID-related symptoms,** or are otherwise feeling unwell, do not attend in-person classes and contact your health care provider and/or student health immediately. COVID symptoms include: fever greater than 100.4 or chills; cough, shortness of breath or difficulty breathing; fatigue; unexplained muscle or body aches; headache; new loss of taste or smell; sore throat; congestion or runny nose; nausea or vomiting; diarrhea.
- We will all wear **face coverings while in the classroom**, unless you have a documented exemption due to a disability or medical condition.
- We will maintain a **6-foot distance from each other at all times** (except in specific lab/studio courses with other specific guidelines for social distancing).
- We will enter the classroom and **fill the room starting at the front, filing all the way across a row**. When class ends, we will exit the row nearest to the door first; the instructor or TA will give the signal for the next row to exit, in the same manner.
- In any small section or lab class that requires them, **additional measures will be listed in the syllabus and be mandatory for class participation**.
- Online office hours will be available for all students.

Revised August 21, 2020

• This course may be recorded for the sole purpose of sharing the recording with students who can't attend class. The instructor will take care not to disclose personally identifiable information from the student education records during the recorded lesson.
Compliance with these guidelines is required for all; anyone who fails to comply will be subject to the accountability process, as stated in the University's Collected Rules and Regulations, Chapter 200 Student Code of Conduct.

If an instructor has concerns about how a student is following COVID-19 policies and protocols, please report those concerns to the Office of the Dean of Students. You can fill out a COVID Safety Measures Reporting Form here: https://cm.maxient.com/reportingform.php?UnivofMissouriSystem&layout_id=38
By taking the above measures, we are supporting your health and that of the whole Mizzou community. Thank you in advance for joining me and your peers in adhering to these safety measures.

Recordings

University of Missouri System Executive Order No. 38 lays out principles regarding the sanctity of classroom discussions at the university. The policy is described fully in Section 200.015 of the Collected Rules and Regulations. In this class, students may make audio or video recordings of course activity unless specifically prohibited by the faculty member. However, the redistribution of audio or video recordings of statements or comments from the course to individuals who are not students in the course is prohibited without the express permission of the faculty member and of any students who are recorded. Students found to have violated this policy are subject to discipline in accordance with provisions of section 200.020 of the Collected Rules and Regulations of the University of Missouri pertaining to student conduct matters.

Revised August 21, 2020

Course Schedule

The Course Schedule is subject to modification. Please check Announcements and emails regularly to ensure you do not miss any changes. Below is the tentative detailed schedule.

8/24: Introduction
8/26: Linear Regression Review

8/31: Classification Review
9/2: Model Assessment and Resampling

9/9: Linear Regression: Model Selection and Regularization (Shrinkage)

9/14: Linear Regression: Model Selection and Regularization (Dimension Reduction)
9/16: Basis Functions and Splines

9/21: Kernel and Local Polynomial Regression
9/23: Generalized Additive Models

9/28: Tree Based Models: CART
9/30: Tree Based Models: MARS and Bagging

10/5: Tree Based Models: Boosting
10/7: Tree Based Models: Random Forests

10/12: Neural Networks
10/14: Deep Neural Networks I

10/19: Deep Neural Networks II
10/21: Gaussian Process Regression

10/26: Support Vector Machines I
10/28: Support Vector Machines II

11/2: Generalizing Linear Discriminant Analysis
11/4: Cluster Analysis (move to DA II in the future)

11/9: PCA
11/11: Multidimensional scaling and nonlinear dimension reduction

11/16: Hidden Markov models I
11/18: Hidden Markov models II

11/30: Latent Variable Models
12/2: Latent Dirichlet Allocation and Topic Modeling

12/7: General measures for predictive importance
12/9: open

Revised August 21, 2020