

MAT 300: Introduction and What to Expect

August 3, 2024

Table of contents

Syllabus	1
Brightspace	2
Day-to-Day Activities	2
The Big Picture	3
Closing...	4

Syllabus

Major Highlights from the Syllabus: I'll ask you to read the syllabus, but here are the most important items:

- Instructor: Dr. Adam Gilbert
 - e-mail address: a.gilbert1@snhu.edu
 - Office: Robert Frost Hall, Room 311
 - Office Hours: Tuesdays 9am - 11am, Thursdays 2pm - 3pm, Fridays 10am - noon
- Books and software are all freely available, see the syllabus for links.
- Grading Scheme:

Grade Item	Value
Participation	10%
Homework (~3)	10%
Competition Assignments (~6)	40%
Project	40%

- The tentative course outline is truly tentative. Use this as a loose agenda, but expect that it will change as we go more quickly or slowly through particular topics.

About the Competition: The best way to learn regression is by doing regression. I've created a Kaggle competition that is private for students in our class. You'll be in a friendly competition with your classmates this semester. Your task is to predict the time between order and delivery of a *Door Dash* order. Your goal is to *construct the best performing predictive model* (the scoring metric used is root mean square prediction error on the evaluation set with hidden delivery times).

You'll need to create a free [Kaggle](#) account before gaining access to the competition. Once you are logged into your Kaggle account, you can access the competition information and data using the link posted to the `#competition-discussion` channel of our Slack Workgroup. I hope that you'll find this friendly competition to be fun and motivating.

Project: Students in MAT300 will complete a large-scale project at the end of the semester. That project will include a full *analytics report*. We'll discuss this in greater detail towards the final month of our semester.

Brightspace

Brightspace will contain much of our course material. The following sections will be used:

- **Syllabus:** Our class syllabus will always be available here
- **Notes:** This section of our BrightSpace course will simply link you back to the webpage I built for this course.
- **Datasets:** The notes that I've prepared for use in our class use a fairly famous, but small datasets on *penguins* and another on *home sales*. While these datasets are really convenient to use for showing you the course material, it will be beneficial to try applying the techniques introduced on your own to some new data. I'll typically use a dataset on San Francisco Apartment rentals in class, but I invite you to choose something you are interested in – I'm also happy to work with different datasets that the class shows interest in. I'll post some links to data sets in this section of BrightSpace.
- **Homework:** I'll post homework assignments and solutions here.
- **Projects:** Generic information about the projects will go here.

Day-to-Day Activities

Here's what you can expect to see in class:

- **Discussion Days** - This course will be infinitely better if we all contribute ideas. I've prepared a complete(*) and very detailed set of notes for you ahead of time, but I'd like to spend class time working discussing and working through the completed analysis

on a *different* data set. I'll provide you a skeleton for taking your own class notes and conducting analysis in Quarto – you can find it on the course webpage and it is also available [here](#). The way you utilize the prepared notes is up to your own discretion.

- **Workshop / Lab Days** - We'll use some of our class meetings to slow down, work with a new data set, and make sure that we are comfortable implementing techniques learned in class with this new data.
- **Lecture Days** - I honestly hope to have none of these. I think that most of the material can be delivered and uncovered via activities and our discussions. I'll spend time explaining the more technical portions of the course material as needed though.

As with every one of my courses, I am absolutely open to changing the structure of course delivery. If the structure of this course isn't working for you, please reach out to me and we can make changes. My goal in this course is for all of you to learn as much about regression and statistical modeling as possible – we can't achieve that if you don't feel like you are benefiting from our class meetings.

The Big Picture

We'll be discussing a lot of material in MAT 300. Here is a very generic road map of what we will discuss. Starting now.

- What are we doing?
 - Data science?
 - Machine learning?
 - Statistical learning?
- Background for working with data
 - What is data?
 - How to work with data: Don't spend it all in one place!
 - How do we visualize and story-tell with data?
 - What if my data is messy? (Spoiler Alert: It will be!)
- What do I need from MAT 240?
- Regression
 - What is regression?
 - Simple regression
 - Assessing model performance
 - The bias-variance trade off
 - Accuracy versus interpretability
 - Multiple linear regression

- How do we deal with categorical predictors?
- Advanced topics
 - What if my dataset has lots of features?
 - Can we penalize models to avoid overfitting?
 - What happens when relationships change?
 - Are there other classes of regression model?
 - Can we build models that predict a categorical outcome?

Closing...

Homework: Please do the following for homework before our next meeting:

- Read Chapter 1 (pages 1 - 14) of the [Introduction to Statistical Learning \(ISLR\)](#) book, or watch the corresponding videos from the textbook authors.
 - The chapter discusses three arenas where statistical learning is applied (regression, classification, and unsupervised learning). We will only discuss regression and a tiny snippet of classification in this course, but reading those example scenarios will give you a better grasp on what it is we are doing in this class.

Next time: we will discuss...

- What is statistical learning in terms of regression?
- Why try to build models (estimate f)?
- What are noise, reducible error, and irreducible error?
- Why are prediction and interpretation competing objectives?
- What are parametric and non-parametric models?
- How do I identify regression versus classification problems?
- What is the difference between supervised and unsupervised learning?