Advanced Methods for Statistical Inference
ITAO70200

Office: 337 Mendoza

Email: seth.berry@nd.edu

Office Hours

Tuesdays & Thursdays – 1:00 to 3:00

These are the *official* office hours. If you find my door ajar to any degree (it will typically be less than 10° and the office will be dark), then you are more than welcome to drop in chat with me about anything (stats, programming, career, etc.).

Class Days and Time

Monday and Wednesday – 3:00 to 4:50

Location – L068

Stat Stroll

Starting this semester, I will be having a weekly *Stat Stroll*. This will involve a 1 to 1.5 mile stroll around campus with a practioneer of the statistical arts. More information will be forthcoming.

Course Objectives

By the end of this class, I want for you to be able use the general linear model, generalized linear models, and interpret results from both. While mastery is not required, you will have a high-level understanding of these methods and you should be able to appropriately apply them to problems.

Attendance

Attendance in this course is not required, in that I will not be taking attendance; attendance is certainly recommended (and even encouraged). While the lecture presentations will be available, they are not verbatim recitations of what was covered and you may not rely on them to make it through everything in the course. Learning statistics takes effort and attending class is but one small part of that effort.

Although attendance is not required, we will have a weekly short comprehension check question. Each question is worth 5 points and is essentially a participation credit. These are meant to be done in class – if you are not in class when they are given, you cannot submit them.

Readings

There is no official textbook for this course, but there are going to be a few assorted readings and resources for topics. I will also share resources for topics. We (read: you) are absolutely not committed to the readings; however, they will give you some very helpful background and I encourage you to read them; in other words, they are suggested readings, not required readings.

Homework

All homework assignments must be submitted in an html file knitted from Rmarkdown. The reasons for this requirement are mainly related to making sure that you have successfully gotten results from your code and that you are not just submitting code without running it first. Once you start using RMarkdown, it will be hard for you to go back to writing in a traditional word processor.

A significant portion of your grade will come from the homework map (i.e., your path towards statistical enlightenment). The skills roadmap has different levels (bronze, silver, and gold). Each level must be completed in order (i.e., you cannot skip silver and go straight to gold from bronze). Silver and gold levels essentially act as modifiers to not only your grade, but also your understanding.

Please feel free to work together on the bronze level of homework (i.e., not the optional sections), but each assignment needs to be your own work. Putting your heads together to formulate an analytic attack plan is perfect (we all stands on the shoulders of giants), but copying and pasting text from each other is unacceptable. In other words, your code and words should not look like one of your classmate's.

Skills Map

The skills map is intended to let you get practice in areas that most interest you personally (maybe categorical data does not really interest you, but you enjoy the GLM). If you really want to dig into the theoretical aspects of the topic, you can! Maybe you want to get a feel for how this might work with “dirty” data. Think of the added levels of the homework as sidequests.

Exam

There will be one exam during the course of this semester and it will be a takehome test. You **cannot** work together on this. You will have one week to complete the test.

Presentations

I promise you that these will be the easiest presentations you will ever give. Select your coursework that you liked best and prepare a 2 minute overview of what you did and what you found (bonus points if you can explain it with one visualization!). These will be given during the designated final day and time. These will be graded on the content of your presentation, not necessarily on the scientific merit of your work! I really want you to be able to speak about your results.

R

There are many statistical programs available and you have likely had some exposure to many of them. In this course, we will be using R exclusively. R is a free and open-source statistical computing language and it is *lingua franca* for modern statistics. We are going to dedicate some time to learning R and working through examples together; our arrangment will be very similar to the science labs that we all remember from our undergrad days. If you have never done any object-oriented programming, it will take a little work – I am only ever an email or visit away.

Grade Breakdown

Comprehension Checks – 30 points (11%)

Homework – 90 points (31%)

Test – 120 points (41%)

Presentation – 50 points (17%)

Total – 290 points

A = 276+ points

A- = 261-275 points

B+ = 252-260 points

B = 244-251 points

B- = 232-250 points

C+ = 223-249 points

Topic 1 – Breaking The Point-And-Click Chains

Wed., 08-22 & Mon., 08-27

R and RStudio

Reproducible Research

Probability, Data Types, Distributions, and Samples

Model Selection

Topic 2 – A Picture Is Worth ~1000 Words

Wed., 08-29

Data Exploration

Data Visualization

Visual Inference

Topic 3 – Categorical Data Analysis

Mon., 09-03

Homework #1 assigned on Mon., 09-03

Visualization

Association

Contingency Tables and χ^2

Correspondence Analysis

Linear Discriminant Analysis

Topic 4 – The General Linear Model

Wed., 09-05 & Mon., 09-10

Homework #2 assigned on Mon., 09-10

Linear regression

T-tests

ANOVA and its variations

Topic 5 – Generalized Linear Models

Wed., 09-12 & Mon., 09-17

Families and Link Functions

Logistic regression

Poisson regression

Mon., 09-17 – TEST 1 ASSIGNED

Topic 6 – Mixed Models

Homework #3 assigned on Mon., 09-24

Wed., 09-19 & Mon., 09-24

Fixed and random effects

Topic 7 – Students' Choice

Wed., 09-26

Regression's Wacky Variants

Robust Models

Quantile Regression

Feature Selection (Ridge -> LASSO -> Elastic Net)

Non-linear Models

Non-linear models

Generalized Additive Models

The Great Statistics Civil War

Bayesian Data Analysis

Words As Numbers

Sentiment Analysis Topic Models

Presentations

Monday, October 01

Location: TBD