**Master of Science in Analytics**

**Advanced Statistics and Modeling**

**Fri 15 Sep, Thu 19 Oct, Fri 20 Oct | 8:00 am – 5:15 pm**

Instructor: Jason Freels

Contact: E-Address: [**jason.freels@afit.edu**](mailto:jason.freels@afit.edu)

Phone: (937) 255-3636 x4676 (office)

Phone: (937) 430-6619 (cell)

Office hours: Mon/Tue 4:00p – 5:00p

These are designated office hours in which I guarantee that I can be available. Feel free contact me at any time - and I’ll get back to you as soon as I can.

**Course Overview**

In the interdisciplinary field of analytics, a good understanding of statistical theory is an important step in being able to draw insights from data. However, many modern analytics problems involve the application of advanced statistical methods and cannot be performed without the use of software. Therefore, this course presents the theory behind various statistical methods in combination with the programming skills required to implement them on data. Topics to be discussed include:

* Methods of producing and collecting data – and what impact that could have on the conclusions derived from an analysis
* Understanding how to manipulate, summarize, and visualize data
* Selecting appropriate methods to analyze data – to detect trends, to check for effects, to make predictions and decisions
* Constructing statistical models to estimate and predict real-world phenomena
* Communicating results in a reproducible manner

**Course Pre-requisites & Co-requisites:**

There are no formal course pre-requisites, but a working knowledge of the fundamental concepts and methods of statistics is presumed. However, because each student brings a different level of statistics expertise to this course, I’ll infuse the course with overview and review of the most important intro stat concepts and methods that are most relevant to statistical modeling and data-based analytics.

There are no formal course co-requisites, but the content presented in this course will overlap in spirit and content with several other courses in the program: Data Mining; Descriptive, Predictive and Prescriptive Analytics; Data Visualization; and Data Management.

**Course learning objectives:**

Reinforce fundamental introductory statistical concepts and methods:

* Connection between data collection and scope of conclusions
* Common visual and numerical summaries, and what to look for in such summaries
* Sampling distributions
* Confidence intervals and significance tests
* Limitations and abuses of formal inference

Understand the statistical modeling process:

* Choosing appropriate models
* Fitting models with technology
* Assessing model fit
* Checking conditions for reliable use of models
* Using and applying models in context

**Assessment:**

Comprehensive assessment will take the form of a modeling portfolio, in which students will be given a problem to investigate and will be required design appropriate data-based studies, collect or acquire the relevant data, carry out the modeling and analysis, and present the results. Additional assessments will be made throughout the course based on each student’s progress toward completing the required assignment post in the DataCamp classroom. Attendance at and participation in all class sessions will be essential parts of the course grade as well.

**Course Format, Resources & Software:**

The in-class portion of this course will mainly be comprised of lectures mixed with coding demonstrations using R. However, with only three days of actual class time, there’s not enough time to get through either a traditional statistics modeling text or a data analytics programming boot camp – let alone both together. To make this possible, I worked with the team at [DataCamp](https://www.datacamp.com) to create a group for this class that will provide each of you with free access to their premium training content for up to six months. Using DataCamp, I can build a curriculum based on their available training courses, post assignments and track your progress. Additionally, the following resources are freely available and may be helpful:

* AFIT Data Science Lab R Programming Guide, <https://afit-r.github.io/>
* Hadley Wickham, Advanced R, 2014, <http://adv-r.had.co.nz/>
* Brian Caffo, Regression Models for Data Science in R, 2015
* Basics of rmarkdown, <http://rmarkdown.rstudio.com/lesson-1.html>

**Course schedule (tentative):**

Fri 15 Sep

a.m. getting started with R

common visual and numerical summaries

uncertainty, sampling distributions

p.m. loss functions, confidence intervals,

significance tests, linear regression

Thu 19 Oct

a.m. project presentations

peer critiques

p.m. ANOVA modeling

logistic regression modeling

Fri 20 Oct

a.m. review of projects

p.m. advanced modelling with h2o and Spark

**Modeling Portfolio project (Fantasy Football)**

Fantasy football is a game in which fans of American football select players from current rosters of teams in the National Football League and area awarded points based on their players performances. The basic idea is to predict which (real-life) players will perform well in an upcoming game. For this project, students will be separated into teams and will be tasked with building a statistical model to predict a player’s performance.

The game which will be the target of this project is the between the Kansas City Chiefs and the Oakland Raiders. Why this game? Because this game will be played on Thursday, 19 October which is the next time we will meet in class. Teams will have until 18 October to build their models and will present their approach in class on the 19th. On the 20th, we’ll compare each model’s prediction against the game results.

**The Rules**

Teams must build a statistical model to predict the number of points awarded to either quarterback using the simplified scoring rules below:

* 1 point is **awarded** for every 30 passing yards
* 6 points are **awarded** for every touchdown (passing, rushing, or receiving)
* 2 points are **deducted** for every interception thrown

Teams must then present their modelling approach to the class using an rmarkdown slide format, teams must discuss: Where the data was obtained, how the data was accessed, their chosen statistical modelling approach, the loss function used by the modelling approach, the factors chosen to include in their model.

Teams will be assessed on the quality of their presentations and the value of their predictions.