MBAC 611 – Advanced Data Analysis - Syllabus

Foote: Fall 2021 (revised 2021-11-18)

Contents

Premise and manifesto

The premise of this course is that *learning is inference*. Learning can be reading, understanding, reflecting whether in our heads or with complex computing environments. We begin with the following chain of reasoning:

- All events, and data collected from events, have a truth value.
- Probability is the strength of plausibility of a truth value.
- Inference is a process of attaining justified true belief, otherwise called knowledge; learning is inference.
- Justification derives from strength of plausibility, that is, the probability distribution of a hypothesis conditional on the data and any background, prior, and assumptive knowledge.
- The amount of surprise, or informativeness, of the probability distribution of data given our experiences, is the criterion for statistical decision making – it is the divergence between what we known to be true and what we find out to be true.

All statistical analysis, and reasoning within analysis, begins from a disturbance in the status quo. The disturbance is the outlier, the error, the lack of understanding, the inattentiveness to experience, the irrationality of actions that is the inconsistency of knowledge and action based on knowledge.

We are surprised when the divergence between what we used to know and what we come to know is wider than we expected, that is, believed. The analysis of surprise is the core tool of this course. In a state of surprise we achieve insight, the *aha!* moment of discovery, the *eureka* of innovation.

The course will boil down to the statistics (minimum, maximum, mean, quantiles, deviations, skewness, kurtosis) and the probability that the evidence we have to support any proposition(s) we claim.

The evidence is the strength (for example in decibels, base 10) of our hypothesis or claim. The measure of evidence is the measure of surprise and its complement informativeness of the data, current and underlying, inherent in the claim.

Learning Goals

At the end of this course students can expect to demonstrate the following goals.

1. Pose a researched business question, model the causal influences implicit in the question, and simulate potential causal relationships and counterfactual inferences.
2. Deploy analyses which produce interactive analytical products using an industry-grade computational platform.
3. Using distributional analysis summarize experience and beliefs about data and using multi-level linear and non-linear models analyze the processes that generated data used to infer potential outcomes to answer business questions.
4. Practice quantitative critical thinking skills through statistical problem solving.
5. Understand the role of the analyst and the analytics process in the decision-making context of complex organizations and their environments.
6. Communicate analytical results to consumers of analytical products effectively using tables and graphs.

Learning outcomes from this course are strongly coupled with the programming goals of the School of Business:

1. Gain experience and expertise in analytical decision making
2. Develop an understanding of leadership
3. Demonstrate an understanding of ethical issues in business
4. Demonstrate an understanding of organizations and the competitive environment

This course will support the attainment of these goals through various group and individual activities throughout our time together this semester. Assignments and other evidence of your work and performance in this course directly align with these goals.

Readings, R Access, Data, and Tutorials

Readings

The main resources for the course are these books:

1. James D. Long and Paul Teetor. 2019. *R Cookbook 2nd Edition*. O'Reilley: Sebastopol, CA. A version of this resource is accessible here with extensive R and R Studio installation instructions.
2. William G. Foote. 2020. *Probabilistic Reasoning*. Access here. This book compiles much of the material needed for the course from other sources and on its own. It is inspired by E.T. Jaynes *Theory of Probability: The Logic of Science* and a pre-print copy is in this directory. and the next resource by Richard McElreath.
3. Richard McElreath. 2020. *Statistical Rethinking: A Bayesian Course with examples in R and Stan* details through this site

Here are additional resources.

3. Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*, Springer Science & Business Media, 2009. The authors have a tutorial site with a downloadable edition of this book, R scripts and other materials accessible here. These authors are at the tip of the machine learning spear in the frequentist tradition.
4. Yanchang Zhao, *Data Mining with R.* 2020, online with books, slides, datasets, R scripts, numerous examples here, all from the frequentist tradition.

The weekly live sessions will expand on key aspects of chapters from Foote and McElreath, and others as needed, and prepare us for weekly assignments. R scripts, `RMarkdown` source files, and data sets accompany each week.

R Access

Access R in several possible ways:

1. With your sign-on credentials use Manhattan College remote desktop (DLS 210, 309 or 314) to use R Studio, the integrated development environment (IDE) we will use.
2. Use the RStudio cloud platform
3. Click here to download a document with details about installing R and preparing your computer for R analytics and see as well as Long and Teetor (2019, chapter 1)

Note: Chrome users may need to right-click and select `Save Link As` to download files.

Data

Most of the data for the course comes from R packages such as `rethinking`.

1. Set up a working directory on your computer by creating an R project in RStudio. Typically this is located in the user's documents directory. With this working directory go to RStudio and create a new project to you will save your files.
2. Within the working directory, you can set up a data directory called `data`. This is a sub-directory of your working directory.

When you set up an `R Project` in a directory, forever more that directory will be your working directory, at least for the work being done in that directory. Choose other directories to set up other projects.

Statistics tutorials

The practice of business analytics benefits greatly from advances in statistics and operations research. Here is a statistics primer that we can use to refresh our understanding and use of basic concepts and models often deployed in business analytics. You can access the Statistical Thinking site here

Weekly Schedule

- Week 1: Basic probability, R, RStudio, rstan, binomial and Poisson applications, grid and approximation methods of computation. McElreath 1-3
- Week 2: Linear regression and its many difficulties including multicolinearity and heteroskedascity. McElreath 4, 5, 6.1 (collinearity).
- Week 3: More modeling issues mainly of the confounding nature; an excursion into the criteria for model selection. McElreath 6, 7.
- Week 4: Interactions; the generalized linear model. McElreath 8, 10, with just a little bit of computational umph from chapter 9.
- Week 5: Logistic and Poisson regression, ordered and unordered categorical variables. McElreath 11, 12.

- Week 6: Multi-level memory, finally varying the slopes along with the intercepts. McElreath 13, 14.
- Week 7: The really important issue of measurement error, missing data; a peak at time series with a twist. McElreath 15, 16.

Yes a lot! But remember we will gloss over some topics and drill into others.

Each week we will practice on short problems to build skills. I will host and record live sessions to go over solutions and help us all shoot trouble in R and in modeling. Each week will have a longer problem to solve called a case. There will 7 such practice sets. You may use your own data with my consultation to develop your own case studies. The skill builders are all or nothing grades. The cases may be revised and resubmitted for re-grading until the day that grades are due to the registrar.

The final grade will be composed of 30% participation in discussion boards and viewing videos, 30% skill-builders, 40% practice sets.

Additional matters

Audience

This course is designed for graduate students interested in building business analytical solutions to business problems using a decision-centric approach. It is offered as an elective in the Business Analytics concentration in the MBA program, and can also be taken by graduate students from other programs in order to complete their course requirements.

Course credits

The successful completion of this course will earn the student 3 credit-hours.

Pre-requisites

While the catalog description and syllabus do not list pre-requisites, we strongly urge students to complete R for MBA students and Business Analytics or possess equivalent knowledge and practice in programming.In any case, the first week will be a crash course in basic R. We will be using several templates and reuse them. McElreath and Foote will have nearly all of the R you will need to consume.

Grading

Grades for work performed in this course are distributed as follows.

1. Seven (6) team cases, equally weighted, constitute 60% of the final grade.
2. Personal participation in the course, measured by completing all of the discussion boards on Moodle and viewing the 6 live session videos, contributes to 20% of the final grade.
3. One (1) personal project constitutes 20% of the final grade.

Students, in teams of two to four, will both be given the same score for a completed project. A final, individually executed and submitted project is due at the end of the term during final exam time.

Grading rubric

All assignments will follow this general rubric:

- **Words:** The text is laid out cleanly, with clear divisions and transitions between sections and sub-sections. The writing itself is well-organized, free of grammatical and other mechanical errors, divided into complete sentences, logically grouped into paragraphs and sections, and easy to follow from the presumed level of knowledge.
- **Numbers:** All numerical results or summaries are reported to suitable precision, and with appropriate measures of uncertainty attached when applicable.

- **Pictures:** All figures and tables shown are relevant to the argument for ultimate conclusions. Figures and tables are easy to read, with informative captions, titles, axis labels and legends, and are placed near the relevant pieces of text.
- **Code:** The code is formatted and organized so that it is easy for others to read and understand. It is indented, commented, and uses meaningful names. It only includes computations which are actually needed to answer the analytical questions, and avoids redundancy. Code borrowed from the notes, from books, or from resources found online is explicitly acknowledged and sourced in the comments. Functions or procedures not directly taken from the notes have accompanying tests which check whether the code does what it is supposed to. All code runs, and the `R Markdown` file `knits` to `html_document` output, or other output agreed with the instructor.
- **Modeling:** Model specifications are described clearly and in appropriate detail. There are clear explanations of how estimating the model helps to answer the analytical questions, and rationales for all modeling choices. If multiple models are compared, they are all clearly described, along with the rationale for considering multiple models, and the reasons for selecting one model over another, or for using multiple models simultaneously.
- **Inference:** The actual estimation and simulation of model parameters or estimated functions is technically correct. All calculations based on estimates are clearly explained, and also technically correct. All estimates or derived quantities are accompanied with appropriate measures of uncertainty.
- **Conclusions:** The substantive, analytical questions are all answered as precisely as the data and the model allow. The chain of reasoning from estimation results about the model, or derived quantities, to substantive conclusions is both clear and convincing. Contingent answers (for example, "if X, then Y, but if A, then B, else C") are likewise described as warranted by the model and data. If uncertainties in the data and model mean the answers to some questions must be imprecise, this too is reflected in the conclusions.
- **Sources:** All sources used, whether in conversation, print, online, or otherwise, are listed and acknowledged where they used in code, words, pictures, and any other components of the analysis.

Grading scale

Thus the numerical grades evolve into letter grades.

A	93% and above
A-	90% to 92.9%
B+	87% to 89.9%
B	83% to 86.9%
B-	80% to 82.9%
C+	77% to 79.9%
C	73% to 76.9%
C-	70% to 72.9%
D+	65% to 69.9%
D	60% to 64.9%
F	Below 60%

Assignment Formatting

All assignments must be turned in electronically, through the learning management system, by each student. All assignments will involve writing a combination of code and actual prose. You must submit your assignment in a format which allows for the combination of the two, and the automatic execution of all your code. The easiest way to do this is to use `R Markdown`. `R Markdown` also allows the use of interactive modeling through `Shiny` applications.

Work submitted as Word files, unformatted plain text, etc., are not acceptable at any time during the course. Each assignment will require the submission of at least one `R Markdown` script file and the `html` file that the

`R Markdown` script generates. When using data sets, this course will only use `csv` (comma separated variable files generated by Excel or in text files or calls to data using APIs. If the submission uses a `csv` file, that file must also be submitted with the `R Markdown` script and generated `html` output files. The student may also submit a supplemental `R` script file, suitably commented, that represents the `R` code chunks in the `R Markdown` script.

Managing the data base of submitted assignments throughout the course will be aided by standards including file name construction for assignment submission. To this end, every file submitted must have a file name which includes the student's name, course identifier, and clearly indicates the type of assignment (project) and its number (week). Here is the format we will use: `yourName_courseidentifier_Assignment#.ext`, where # is the week number and ext is the file name extension. For example W.G. Foote would submit an `RMarkdown` file with this filename: `wgfoote_MBAC611_Assignment1.Rmd`, where the file extension `Rmd` is the extension that `RStudio` uses for `R Markdown` documents. File extensions `R`, `html`, and `csv` are the other three admissible file types.

Course Specific Policies

Students are expected to behave in a professional and courteous manner at all times when interacting with all members of the course learning community. Respect for others is demonstrated through attendance, meaningful participation, and punctuality. Every effort should be made to be present for each session, if not feasible, view the recording of each session, especially since weekly assignments will be made conditional on content in live sessions.

All projects must be completed and submitted by the due dates and times set out. This will allow the entire class to review and revise submissions in a timely fashion. Submissions to `lms.manhattan.edu` are based on eastern (US) time. Late submissions will result in student inability to accumulate the knowledge needed to advance to the next week's coverage of course topics. Late submission will also delay necessary instructor feedback to the student in a timely fashion. As the course continues to layer on more skills and capabilities, a late submission with inaccurate or incorrect implementations of financial applications will only deprecate the student's ability to successfully complete future assignments.

Academic Integrity

Manhattan College's Academic Integrity Policy holds students accountable for the integrity of the work they submit. Students should be familiar with the policy and know that it is their responsibility to learn about course-specific expectations, as well as about policy. The policy governs appropriate citation and use of sources, the integrity of work submitted in exams and assignments, and the veracity of signatures on attendance sheets and other verification of participation in class activities. The policy also prohibits students from submitting the same written work in more than one class without receiving written authorization in advance from both instructors. The standard sanction for a first offense by a graduate student is suspension or expulsion. For more information and the complete policy, see http://academicintegrity.syr.edu/academic-integrity-policy/.

In this course, all sources, whether verbal, online, in print, or other, must be cited following prevailing business and academic requirements and practice.

Student Academic Support Services | Center for Academic Success

The Center for Academic Success (CAS) is committed to providing student-centered programs and initiatives designed to enhance learning and promote success and persistence for all Manhattan College students. Students will work collaboratively with qualified peers and professionals to develop knowledge, skills and strategies needed for success in the classroom and beyond.

The CAS has two locations; the Learning Commons in Thomas Hall 3.10 and the Leo Learning Center in Leo 117/118. Services include online and in-person individual tutoring, online small group peer tutoring (select courses), Supplemental Instruction (select courses), student academic success coaching, and online writing center services. All services are free of charge and available to all Manhattan College students. Appointments are preferred but walk-ins are welcome. To make an appointment, students can log into their Jasper Connect

account or visit the CAS in Thomas Hall, 3.10. Students can also contact success@manhattan.edu with any questions.

For more information about these services please visit the CAS webpage here, and to learn about CAS Fall 2020 return to campus safety efforts please visit the One Manhattan webpage here.