**Statistical Modeling and Decision Making**

**Course Term:** Summer 2020 (August 17, 19, 21, 24, 26, 28; 8:00-11:15 PM EDT)

**Course Format:** Online (Zoom)

**Office Hours:** TBD

**Instructor:** [Elliot Shin Oblander](https://www8.gsb.columbia.edu/cbs-directory/phd/EOblander23) (eoblander23@gsb.columbia.edu)

**TA:** [Jasmine Yang](https://www8.gsb.columbia.edu/cbs-directory/phd/YYang24) (yyang24@gsb.columbia.edu)

“All models are wrong, but some are useful.” -George Box, 1978

**Course Motivation and Objectives:**

Most business problems require making decisions with limited information. That is, we often must make decisions based on inputs that are not perfectly know to us: to pick between several different product designs, we need to understand how people might respond to each design; to decide how much money to spend on acquiring a customer, we need to understand how much revenue that customer might ultimately bring in; in pricing a product, we need to understand how different prices might affect consumers’ short-term purchase patterns and long-term attitudes about the product.

In this course, we will learn about *statistical modeling* as an approach to solving these problems. Statistical modeling allows us to express what we do and don’t know about a problem, to estimate and predict unknown quantities, and to use those estimates to make informed decisions.

By the end of the course, you should:

* Know how to translate business problems into statistical problems*.*
* Have a basic toolbox of models and methods for solving statistical problems, and be able to identify the “right tool for the job” for a given statistical problem.
* Understand the assumptions required to apply these methods to a business problem of interest and be able to assess the validity of these assumptions.
* Know how to translate statistical insights back into business insights.

The overarching goal of this course is to equip you with the intuition and toolbox needed to conduct basic statistical modeling and use models to answer practical questions. This will help prepare you for more advanced courses where you will learn about rigorous statistical inference (B9106: Applied Multivariate Statistics) and advanced modeling techniques (B8683: Marketing Models), as well as MBA courses where you will be challenged to translate quantitative analyses into managerially relevant insights.

**Prerequisites:**

Students are expected to enter the MS in Marketing program with a working knowledge of:

* **Probability:** discrete and continuous random variables; computation of mean and variance of distributions; joint, marginal, and condition probability distributions.
* **R programming:** programming fundamentals (variable types/classes, loops, functions); data loading and cleaning; basic exploratory data analysis and visualization; basic simulation.
* **Matrix algebra:** matrix addition, multiplication and inversion; linear transformations and projections; eigen decompositions.

We will primarily make use of probability and R programming in this course, with elements of matrix algebra as needed. Matrix algebra will primarily be important for the Multivariate Statistics and Marketing Models courses in the fall term.

**Course Materials:**

There is no required textbook for this course, but the book “An Introduction to Statistical Learning” ([available for free online](http://faculty.marshall.usc.edu/gareth-james/ISL/)) may be a useful resource.

I will post lecture notes and recordings on Canvas, and will provide references on where you can learn more about specific topics if you are interested.

**Course Requirements and Grading**

Evaluation will be based on the following components and weights:

*Prerequisite learning check (15%)*

* I will post 3 short assignments on Canvas that will help you get practice in each of the course prerequisites (probability, R, and matrix algebra). I will also post online resources that will help you get up to speed if you are not familiar with any of these areas.
* These assignments will be graded mostly by completion, and we will review the more difficult questions as needed in the first couple of class sessions.

*Class participation (15%)*

* I will maintain a Piazza page for the course where you can ask questions about course-related material (e.g. questions about lecture material or the homework), answer each other’s questions, or post interesting relevant material (e.g. blog posts or news articles). I encourage you to participate in this platform and learn from each other.
* For those of you attending the class in real time, I will solicit class participation regularly and expect you to ask questions and raise concerns when you have them. You should also post on Piazza (and/or attend office hours if you are in a time zone that you can make it) to discuss class material, homework, etc.
* For those of you unable to attend in real time, I expect you to watch the lecture recording as soon as possible (e.g. the morning after the actual class occurrence) and post questions and comments on Piazza and/or during office hours based on the lecture recording.

*Regular post-class assignments (70%)*

* After each class, I will post an assignment to review and reinforce the material we covered in that day’s class. Given the compact structure of our course, these assignments will be fairly guided. These assignments will involve conceptual questions, derivations, and applications of models to data.
* The assignments are due within 72 hours of the end of class.
* You may discuss the questions with classmates, but *all work must be your own*: you must write your own code, do your own derivations, write your own answers, etc. If you discussed the assignment with other classmates, *say who you discussed with* at the top of the assignment you turn it.

**Course Topics and Schedule**

**Session 1 (Monday August 17): Overview of Course; Review of Probability Theory**

Conceptual overview of statistics, probability, and modeling as they relate to decision making and business. Review of univariate random variables, expectations, variances, and transformations. Overview of some useful probability distributions and the relationships between them.

**Session 2 (Wednesday August 19): Models for Multiple Variables**

Joint distributions of multiple variables. Covariance and correlation. Bayes’ theorem, conditional distributions, and posterior inference. Special cases of mixture models, factor analysis, and regression.

**Session 3 (Friday August 21): Taking Models to Data, Part 1**

Using sample analogs to estimate statistics of distributions. Estimating a statistical model: maximum likelihood estimation as a versatile tool for taking models to data. Computing standard errors and confidence intervals. Diagnostics for model checking and model comparison. (Time permitting: empirical Bayes and alternate methods for model estimation.)

**Session 4 (Monday August 24): Taking Models to Data, Part 2**

A probabilistic perspective on regression and least squares estimation. Issues of interpretation, collinearity, and identification. A basic overview of generalized linear models. Model robustness and misspecification: implications for inference and interpretation. (Time permitting: a basic introduction to machine learning models)

**Session 5 (Wednesday August 26): Translating Business Problems to Statistical Problems**

Introduction to statistical decision theory; understanding common statistical procedures as decision problems. Framing business problems as statistical decision problems; application to online advertising and A/B testing.

**Session 6 (Friday August 28): Correlation vs. Causation**

The bridge from statistical models to causal models. Endogeneity/confounding, randomized experiments, and the assumptions necessary to estimate causal effects. Illustrations in the context of pricing and online advertising.