# Live Session 1 - Discrete Response Models

*Jeffrey Yau*

*September 5, 2017*

## Agenda

1. Introdution (30 minutes, depending on the number of students attending the sessions)
2. A Discussion of weekly workflow and tips for success in this course (20 minutes)
3. An overivew of topics covere in this lecture (5 minutes)
4. Discussion of the analysis of two binary variables (35 minutes)

---

## 1. Introduction (30 minutes, depending on the number of students attending the sessions)

1. Instructor's self introduction
2. Students' self introduction: each student takes turn introducing himself/herself (3 minutes each), addressing the questions below
   - Did you take the new or old version of *w203*?
   - Which is your cohort?
   - What company are you working for, and what's your role?
   - Do you use machine learning or statistic modeling in your current work? If so, what techniques do you use?
   - Why do you take this course?
3. Course Overview, Other Reminders, Q&A

---

## 2a. Quick Introduction for this course (5 minutes)

- Professor Yau to give an overivew of this course, addressing various perspectives though this course primarily focuses on statistical models for various types of response variales:
  1. the traditional statistical perspective
  2. the modern statistical perspective
  3. machine learning perspective
- This course focuses on statistical model building from the modern perspective used in data science, covering steps from defining a (business, policy, scientific, etc) problem that can be addressed using data at hand, conducting EDA as a pre-model-building step, understading the underlying statistical assumptions of a model under consideration, specifying a model, engineering features, estimating a model using functions in R, testing hypothesis, evaluating a model, conducting model diagnostics, and testing model assumptions. On occasion basis, I will also compare models covered in this course with machine learning techniques that can be used to solve similar problems.

## 2b. Weekly Workflow (5 minutes)

A typical week of the course proceeds as follows: - Before live session: Watch all async content, study the assigned readings, attempt some of the end-of-chapter exercises, and write down questions / comments that you want to discuss in the live session - In live session: Please come to the live session prepared. **Live sessions are not lectures.** We will review some of the more-difficult-to-understand concepts, build upon the async to extend your understanding, occasionally provide additional examples, and introduce additional exercises. To encourage active participation, we will ask individual student to explain certain concepts covered in the async lecture - After live session: Review the materials covered in the previous week and continue to attempt some of the end-of-chapter exercises. There are numerous exercises at the end of each chapter. Do as many of them as you can.

## 2c. Professors' Expectation and How to Succeed in this Class (10 minutes)

**Please take out the syllabus and review some of the highlights together**

Here are some strategies based on our past experience and how I designed the course

- Review materials taught in the new version of w203, especially the part on linear regression modeling.
- **On watching the async lecture:** Different people learn differently; as such, it is difficult to give a "general" rule of how one should watch the async lectures. That said, I think it is very rare that students can watch each lecture in one sitting and master the materials. It is more likely that you will have to watch a section; pause; read the corresponding sections in the text and work through some examples; and then rewatch the section of the lecture to ensure that you understand the materials. This is why I followed so closely the assigned textbooks, which I chose by considering many different texts. This is also why mastering the concepts and methods in this course is very time consuming.
- Study your readings. More importantly, do the exercises in the texts. This is an advanced course; it is very difficult, if not impossible, to learn statistical modeling by simply reading textbooks or watching videos; doing hands-on work is critical to learn the materials taught in this course (and any statistical modeling and machine learning course). Note that it is also impossible to cover every single concept in a ninety-minute async lecture; you will have to build upon the materials taught in the async lecture by studying the readings and doing many exercises.
- Come to the live session prepared. I expect that you should watch the async video lecture, study the assigned readings, work on the examples in the book, and attempt a at least a few end-of-chapter exercises. In the live session, we may occasionally ask questions that you can only answer if you keep up with the readings.
- Form study groups! Do it now (if you haven't done so already)!
- Do not skip the readings or async lectures. In this course, once you fall behind, it is extremely difficult to catch up, as we cover a different statistical model almost every week.
- For the lab, do not use the **"divide-and-conquer"** strategy if you work in a group, although you are encouraged to discuss your work (after you finish them) with your group.
  - Attend office hours to ask questions related to the concepts and techniques covered in async lectures and readings; don't just come to office hours when you have questions on the labs. **Note that for the labs, we will only answer clarifying questions in the office hours.**
  - Labs: After each of the lab due date, we will distribute the solution; as such no late work will be accepted. We will also return the graded lab to you within two weeks or so with general comments for good and not-so-good practices observed in all the labs. Please talk to me or Professor Tiwari

during office hours if you have questions about the garded lab **after you carefully review our solutions**.

## 3. Topics covered in this lecture

- An introduction to categorical data, Bernoulli probability model, and Binomial probability model
- Computing the probability of binomial probability model
- Simulating a binomial probability model
- Estimating the Binomial probability model using maximum likelihood estimation (MLE)
- Confidence intervals:
  - Wald confidence interval
  - Alternative confidence intervals
- Hypothesis test for the probability of success
- The case of two binary variable
  - Contingency tables
  - The notions of relative risks, odds, and odds ratios
- Two Binary variables
  - Contingency table
  - MLE
  - C.I.s for the difference of two probabilities
  - Relative Risks
  - Odds
  - Odds ratios (OR)
  - log(OR)
  - Estimation and inference

In the live session today, I want to focus on (1) some of the derivations that are missing in the book and MLE, (2) confidence intervals, and (3) the concepts of relative risks and odds ratios. I will leave a few take-home exercises as well.

## 4. Derviations of Binomial Regression Models:

Recall that the probability mass function of the Binomial random variable is

$$P(Y_j = w_j) = \binom{n}{y_j} \pi_j^{y_j} (1 - \pi_j)^{n - y_j}$$

where $y_j = 0, 1, \ldots, n$ where $j = 1, 2$

$$E(Y_j) = \mu_j = n\pi_j$$

$$Var(Y_j) = \sigma_j^2 = n\pi_j(1 - \pi_j)$$

- Refer to the notes *"Binomial Model Derivation"* to be distributed in live session this week