W271: Statistical Methods for Discrete Response, Time-Series, and Panel Data

Master of Information and Data Science

UC Berkeley

Fall 2022

Course Designer and Developer: Dr. Jeffrey Yau

Live Session Instructors:

Dr. Majid Maki
Email: makinay1363@berkeley.edu
Preferred contact: Course Slack Channel
Office Hours: 5:00-6:00 p.m. Tu-Fri

Vinod Bakthavachalam
Email: bvinod@berkeley.edu
Preferred contact: Course Slack Channel
Office Hours: 5:00-6:00 p.m. M-W

Teaching Assistant:

Kyle Chuang
Email: kchuangk@ischool.berkeley.edu
Preferred contact: Course Slack Channel
Office Hours: 6:00-7:00p.m. M

Course description

This course covers a range of statistical techniques to model cross-sectional data with unordered and ordered categorical response, count response, univariate time-series data, multivariate time-series data, longitudinal (or panel) data, and multi-level data from data science perspective. It teaches how to choose from a set of statistical techniques for a given question and to make trade-offs between model complexity, ease of interpreting results, and implementation complexity in real-world applications. It emphasizes on the use of exploratory data analysis (EDA) to generate insights for subsequent statistical modeling as applied to solving data science problems that are often given as (vague) business or policy questions. In addition, it covers the mathematical formulation of the statistical models, assumptions underlying these models, the consequence when one or more of these assumptions are violated, the potential remedies when assumptions are violated, hypothesis testing, model selection, model diagnostic, model assumption testing, and model evaluation. The course goes well beyond the simple mechanical implementation of statistical methods using statistical software, such as R. The design principles of solutions and theoretical foundations of the statistical models that make up the solutions are the major focus, as they are essential for data science practitioners.

Throughout the course, we emphasize formulating, choosing, applying, implementing, evaluating, and testing statistical models to capture key patterns exhibited in data. All of the techniques introduced in this course come with examples using real-world and simulated data, and some come with R codes. As concepts in probability, mathematical statistics, and matrix notations are used extensively, students should feel very comfortable with the definition, manipulation, and application of these concepts in mathematical notations.

Prerequisites

1. Passing DataSci W203 with at least a B+ and having a very solid understanding of the probability and mathematical statistic concepts and techniques and linear regression modeling
2. Hands-on experience in R
3. Working knowledge of calculus and linear algebra
    - Note that differential calculus, integral calculus, matrix notations, and probability concepts are used extensively throughout the course.

Expectations on the Students

The asynchronous video lectures and the assigned textbook readings are mandatory. Students are expected to watch the asynchronous lectures and study the corresponding textbook chapter(s) or article(s) before attending the live sessions, where group exercises are assigned, and inclass discussion are conducted. Attendance and participation in live session are mandatory.

Remarks on asynchronous videos and readings

Remarks on asynchronous videos and readings: As many concepts covered in this course are quite abstract, most students will need to watch the asynchronous videos a few times and perhaps even watch a couple of modules, read the corresponding sections in the assigned readings, and try out a few examples before even moving on to the next video module. It would be rare that one can watch the asynchronous video lecture only once (and in one sitting) and understand all of the concepts and techniques covered in that week. To aid the studying, I designed the course to follow the text very closely in most of the lectures, especially the first five lectures, attempting only to highlight the important concepts and techniques in the asynchronous lectures. I adopted the specific textbook for the discrete-response-model portion (i.e. first 5 lectures) of the course because the authors also provide their own materials and videos.

Remarks on live sessions

Live sessions are not lectures; the live session instructors will not be lecturing during the live sessions. When attending the live sessions, students should find a place with good internet connection. If you mute your video during the live session, the professor will ask that you unmute your video. Students are expected to actively participate in the live session and contribute to the discussions. Students should also come to the live sessions with questions that they would like to discuss with classmates and the instructor. Ideally, the students can post the questions to the course's slack channel wall in advance so that the instructor and other students can think about them before the live session. It is important to note that live sessions are not lectures, though the instructors occasionally may spend some time review key concepts covered in the asynchronous lectures and/or the readings. It is also important to know that the asynchronous video lectures and the assigned textbook readings are not substitutes for each other. Students should also attempt as many end-ofchapter exercises as they can both before and after live sessions. The textbooks go into a lot more details than the asynchronous lectures and provide many more examples that are not possible to cover in a 90-minute asynchronous lecture. Therefore, students are expected to study the readings and will be tested on the mastery of the concepts and techniques covered in the assigned readings.

As mentioned, this is a fast-paced course, and the mathematical structure and assumptions of the statistical models taught are covered in-depth. That said, extensive proofs, derivations of properties of estimators, derivations of standard error of estimators, and the numerical techniques underlying the estimation methods will not be emphasized or even mentioned in most cases. Notions of probability theory and mathematical statistics and matrix algebra are used extensively throughout the course. While we cover the mechanical implementation of these models using computer codes, the course focuses on building statistical models that can be applied to realworld data science problems and goes well beyond the mechanics. In fact, many of the R libraries introduced in this course have more functions than we have the time to cover. Therefore, students are expected to read the documentation associated with these libraries and learn how to apply the functions in the libraries to build statistical models. For these reasons, it is not uncommon for students to spend on average 15 to 20 hours per week studying materials in this course in addition to the time spent watching the asynchronous lectures and attending the live sessions. There are weeks, especially towards the end of the course, that may

This is not designed as a graduate-level mathematical statistics course. This is a statistics course for aspiring or existing data scientists who want to learn some basic statistical techniques to model categorical, time-series, and panel data. This is also not a course to teach statistical modeling used in a specific scientific discipline. This course emphasizes data science applications of statistical techniques. A good understanding of the mathematical underpinnings of the models is critically important to apply these models correctly to solve real-life data science problems. However, heavy emphasis on applications also means that we downplay the mathematical proofs not because they are not important but because (1) it requires a lot more time in both the asynchronous lectures and out-of-classroom self-study time by students and (2) it requires that students be very comfortable with the concepts of stochastic convergence. Therefore, students should expect that this course is designed for aspiring data scientists and not for Ph.D. statisticians, econometricians, or scientists of various disciplines. Some former students who came to this course with the wrong expectations that every single proof of the theorem or derivation be provided or specific techniques be taught and left with disappointment.

Live Session Attendance

While not counting to towards the grade, attendance and active participation during live sessions are strongly mandatory. Students do not need to explain to the professor the reason of missing class. Note that we have a "no-muting-video" policy. Make sure you have good internet connection when attending the live sessions, as I will ask you to unmute your video.

Office Hours

Both the Teacher Assistants and the instructors will have office hours each week. Office hours are designed to discuss materials covered in the asynchronous lectures, assigned readings, and discussions in the live sessions. We encourage you DataSci w271 UC Berkeley Master's in Information and Data Science 5 to join us in the office hours to discuss course related materials. For personal questions, we can address them if there are no other students attending the office hours, or you can make additional appointments with one of us.

Required Textbooks and Other Course Resources

1. **[BL2015]** Christopher R. Bilder and Thomas M. Loughin. *Analysis of Categorical Data with R*. CRC Press. 2015.
2. **[CM2009]** Paul S.P. Cowpertwait and Andrew V. Metcalfe. *Introductory Time Series with R*. Springer. 2009. (ISBN-10: 978-0-387-88697-8)
3. **[TSA]** Robert H. Shumway and David S. Stoffer. *Time Series Analysis and Applications*.
4. **[HA]** Rob J Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*.

   https://otexts.com/fpp3/

   - Note that because this is an online book; the authors may update book after this syllabus was written. So, please ensure that the assigned sections correspond to the topics covered in the corresponding week. When in doubt, please contact the instructors or the TAs.
5. **[W2016]** Jeffrey Wooldridge. *Introductory Econometrics: A Modern Approach. 6th edition*. Cengage Learning. (ISBN-10: 130527010X)
6. **[BMBW]** Douglas Bates, Martin Machler, Benjamin Bolker, and Steve Walker. *Fitting Linear Mixed Effect Models Using lme4*
7. **[JF2016]** John Fox. *APPLIED REGRESSION ANALYSIS and GENERALIZED LINEAR MODELS(Third Edition)*
8. Additional papers, articles, and readings may be provided throughout the course.

Grade Assignment:

Grade	Range
A	93-100
A-	90-92
B+	87-89
B	83-86
B-	80-82
C+	77-79
C	73-76
C-	70-72
Lower Grades	< 70

Assignment weight

Item	Weight
Three Group Assignments (Labs)	45 %
Seven Homework Assignments	35 %
12 Async Review quizzes	15 %
Participation	5 %

Assignments schedule

Date	Unit	Assignment availability	Assignment due
2022-08-23	1	Quiz-1/HW-1/Lab-1	-
2022-08-30	2	Quiz-2/HW-2	Quiz-1/HW-1
2022-09-06	3	Quiz-3/HW-3	Quiz-2/HW-2
2022-09-13	4	Quiz-4	Quiz-3/HW-3
2022-09-20	5	Quiz-5	Quiz-4
2022-09-27	6	Quiz-6/HW-6/Lab-2	Quiz-5/lab-1
2022-10-04	7	Quiz-7/HW-7	Quiz-6/HW-6
2022-10-11	8	Quiz-8/HW-8	Quiz-7/HW-7
2022-10-18	9	Quiz-9	Quiz-8/HW-8
2022-10-25	10	Quiz-10	Quiz-9
2022-11-01	11	Quiz-11/HW-11/Lab-3	Quiz-10/Lab-2
2022-11-15	12	Quiz-12	Quiz-11/HW-11
2022-11-29	13	-	Quiz-12
2022-12-06	14	-	Lab-3

Grading and Feedback Method

Our method of feedback for this course respects you as students who are in charge of your learning. The instructors and TAs are going to provide what is referred to as "mastry-level" feedback. This feedback allows you to quickly understand if you have met the core objectives of the questions, or whether your work requires changes to come up to a level of demonstrated mastry.

We will release solution sets for homework either as soon as all student work has been completed, or two-days after the last deadline of the week. For example, if there is a section that meets on Thursday at 4:00pm, if all students in all sections have turned in their work, we will release solutions at the of the section. If there are students who have fallen behind, we will wait to release solutions until Saturday at 4:00pm. There is much to learn from comparing other's approaches to problem solving, and we would encourage students to consider the review of the solution sets as a core part of the learning for the assignment!

If you have successfully and competently completed all parts of a question, we will mark that question as having been "mastered" and you will receive full marks. If you have shown considerable competency, but there are some small places for your work to improve, it will be marked as "complete" and you will receive 90% of the full marks score. If your attempt is missing core concepts or core pieces of the work to be done, it will be marked as "attempted" but not complete or mastered, and you will receive 50% of the full marks score. And, if you fail to attempt a problem, then it will be marked as "not attempted" and you will receive 0% of the full marks score.

Our goal with this mastry-level marking is to *very* quickly provide you with feedback about your work relative to that of a professional caliber. With this feedback, we expect that you will review the solution set, evaluate where yours and our approaches are similar and different, and continue to learn *even after* the assignment deadline.

Finally, and importantly with this mode of feedback, we hope that students and instructors can have open conversations about aproaches to solving these questions in office hours. After each deliverable, we will dedicate office hours time toward these dicsusions so that we can all continue to learn together.

Late Policy

We set deadlines for homework, Async Review quizzes , and labs to be completed. At the same time, we understand that MIDS is a single facet of the life that you live, and some delays are unavoidable.

- **Homework:** Homework assignments are chances for you to apply the content that you have learned in the async and the live sessions. These occur at a regular (nearly weekly) pace. **Homework is due at the time that your live session meets on the day that it meets.**
  - Students have five "late days" that they can use without penalty on homework through the semester. After those "late days" have been used, each day late will be assessed a 10% penalty from the final grade of that homework assignment. As an example, after using all five late days, a homework assignment turned in one day late could earn a maximum score of a 90% on that homework.
  - Students **cannot turn in any single homework more than two days late**. This balances the need for student flexibility, with the need to release feedback to other students in a section. For example, a student who has class on Monday at 4:00p cannot turn in an assignment later than Wednesday at 4:00p.
  - Note on Gradescope due dates: Gradescope is the online platform that we use for homework. The due dates shown on gradescope are there to accomodate the last section of the week. You still must turn in your assignments before your live session meets. Late days will be counted at the end of the semester. You must keep your own tally of how many late days you have used.
- **Labs:** The labs in the course are group projects. We will expect that labs are turned in for instructor review, in whatever state they are in, at the time they are due.
- **Async Review quizzes:** Quizzes count for a smaller proportion of the final course grade, and happen weekly. **Students have "seven late days for each one" that they can use without penalty**

Statement of Equity, Diversity, and Inclusion

At UC Berkeley, we promote equity, diversity, and inclusion. Our faculty, staff, students, and all other members of our community are accountable for integrating equity, inclusion, and diversity into all aspects of our lives at MIDS.

In an ideal world, science in general and data science in particular would be objective. However, for a variety of reasons, data science could be biased, reflecting a small subset of individuals' behaviors or voices, potentially because of the way data is collected.

In this class, we will make a serious effort to learn statistical models from a diverse group of scientists, statisticians, and econometricians from a set of diverse disciplines. However, it is still possible that bias exists in the materials due to the choice of the examples by the authors of our assigned textbook books, articles, or even code documentation, even though the materials are data science in nature. Integrating a diverse set of experiences is important, beneficial indeed, when learning data science, which is an interdisciplinary subject that uses scientific approaches. Data scientists in industry, academia, government, and other organizations come from a very diverse background, which extends beyond race, ethnicity, country of origin, gender, age, sexuality, religion, and social class: they are trained in different disciplines and speak different technical languages. In the MIDS program, faculty members and students come from a wide range of industries, are trained in many different disciplines, have different numbers of working experiences, spread the whole range of career levels, and possess a wide spectrum of expertise. I find it beneficial, both to the overall outcome of the discussion and to my personal understanding of the subject under discussion, to learn about others' viewpoints when discussing data science topics, and I hope that you will, too. As such, where possible I would like to discuss issues of diversity in data science as part of the course.

Please contact me or submit anonymous feedback if you have any suggestions to adjust the course materials to promote equity, diversity, and inclusion. Furthermore, I would like to create a learning environment for this course that supports a diversity of thoughts, perspectives, experiences, best practices, and honors your identities (including race, gender, class, sexuality, religion, country of origin, ability, experience, etc.). To help accomplish this:

- If you have a name and/or set of pronouns that differ from those that appear in your official records, please let the T.A.s, your classmates, and me know how you would like to be addressed.
- If you feel like your performance in the class is being impacted by your experiences outside of class, please do not hesitate to discuss with Student Affairs or me. I want to be a resource for you. Remember that you can also submit anonymous feedback (which will lead to me making a general announcement to the class, if necessary to address your concerns). If you prefer to speak with someone outside of the course, MIDS Student Affairs would be a good start. The University also has a Division of Equity and Inclusion. More information can be found at https://diversity.berkeley.edu/

  https://diversity.berkeley.edu/

I am myself still learning about diverse perspectives and identities in the context of data science. If something said in class (by anyone, myself included) that made you feel uncomfortable, please talk to me about it. (Again, anonymous feedback is always an option.)

As a participant in course discussions, it is important that you honor the diversity of your classmates and the teaching team.

Course Outline

Lecture 1: Discrete Response Model

- Introduction to categorical data, Bernoulli probability model, and binomial probability model
- Computing probabilities of binomial probability model
- Simulating a binomial probability model
- Maximum likelihood estimation (MLE)
- Wald confidence interval
- Alternative confidence intervals and true confidence level
- Hypothesis tests for the probability of success
- Two binary variables and contingency tables
- Formulation of contingency table and confidence interval of two binary ariables
- The notion of relative risk
- The notion of odd ratios

Readings

- BL2015: Ch. 1
  - Skip Sections 1.2.6 and 1.2.7

Lecture 2: Discrete Response Model

- Introduction to binary response models and linear probability model
- Binomial logistic regression model
- The logit transformation and the logistic curve
- Statistical assumption of binomial logistic regression model
- Parameter estimation
- Variance-Covariance matrix of the estimators
- Hypothesis tests for the binomial logistic regression model parameters
- The notion of deviance
- The notion of odds ratios
- Probability of success and the corresponding confidence intervals
- Visual assessment of the logistic regression model

Readings:

- BL2015: Ch. 2.1, 2.2.1 – 2.2.4
- Additional readings may be assigned

Lecture 3: Discrete Response Model

- Variable transformation: interactions among explanatory variables
- Variable transformation: quadratic term
- Categorical explanatory variables
- Odds ratio in the context of categorical explanatory variables
- Convergence criteria and complete separation
- Generalized Linear Model (GLM)

Readings:

- BL2015: Ch. 2.2.5 – 2.2.7, 2.3
- Additional readings may be assigned

Lecture 4: Discrete Response Model

- Introduction to multinomial probability distribution
- I×J contingency tables and inference procedures
- The notion of independence
- Nominal response model
- Odds ratios
- Contingency table
- Ordinal logistical regression model
- Estimation and statistical inference

Readings:

- BL2015: Ch.3
  - Skip Sections 3.4.3, 3.5
- Additional readings may be assigned

Lecture 5: Discrete Response Model

- Poisson probability model
- Poisson regression model
- Model for mean: log link
- Parameter estimation and statistical inference
- Variable selection
- Model evaluation

Readings

- BL2015: Ch.4.1, 4.2.1 – 4.2.3, 5.2
  - BL2015: Skim sections 5.1, 5.2.3, 5.3, 5.4
- Additional readings may be assigned

Lecture 6: Time Series Analysis

- Introduction to time series analysis
- Basic terminology of time series analysis
- Steps to analyze time series data
- Common empirical time series patterns
- Examples of simple time series models
- Notion and measure of dependency
- Examining time series correlation - autocorrelation function (ACF)
- Notion of stationarity

Readings:

- CM2009: Ch. 1, 2, and 4.2
- HA: Ch. 2

Lecture 7: Time Series Analysis

- Classical Linear Regression Model (CLM) for time series data
  - You will have to review CLM by yourself
- Linear time-trend regression

- Goodness of Fit Measures (for Time Series Models)
- Time-series smoothing techniques
- Exploratory time-series data analysis
- Autocorrelation function of different time series

Readings:

- TSA: Ch. 2
- CM2009: Ch. 5
  - Skip Sections 5.6
- Additional readings may be assigned

Lecture 8: Time Series Analysis

- Autoregressive (AR) models
  - Lag (or backshift) operators
  - Properties of the general AR(p) model
  - Simulation of AR Models
  - Estimation, model diagnostics, model identification, model selection, assumption testing, and statistical inference
- Moving Average (MA) Models
  - Lag (or backshift) operators
  - Mathematical formulation and derivation of key properties
  - Simulation of MA(q) models
  - Estimation, model diagnostics, model identification, model selection, assumption testing, and statistical inference / forecasting

Readings:

- CM2009: Ch. 4.5-4.6, and 6.1-6.4
- Additional readings may be assigned

Lecture 9: Time Series Analysis

- Mixed Autoregressive Moving Average (ARMA) Models
  - Mathematical formulation and derivation of key properties
  - Comparing ARMA models and AR models using simulated series
  - Comparing ARMA models and AR models using an example
- An introduction to non-stationary time series model
- Random walk and integrated processes
- Autoregressive Integrated Moving Average (ARIMA) Models
  - Review the steps to build ARIMA time series model Simulation
  - Modeling with simulated data using the Box-Jenkins approach
  - Estimation, model diagnostics, model identification, model selection, assumption testing, and statistical inference / forecasting, backtesting
- Seasonal ARIMA (SARIMA) Models
  - Mathematical formulation
  - An empirical example
- Putting everything together: ARIMA modeling

Readings:

- CM2009: Ch. 4.3-4.4, 6.5-6.6, and 7.1-7.3
- HA: Ch. 9
- Additional readings may be assigned

Lecture 10: Time Series Analysis

- Regression with multiple trending time series
- Correlation of time series with trends
- Spurious correlation
- Unit-root non-stationarity and Dickey-Fuller Test
- Cointegration
- Multivariate Time Series Models: Vector Autoregressive (VAR) model
  - Estimation, model diagnostics, model identification, model selection, assumption testing, and statistical inference / forecasting, backtesting
  - Notion of cross-correlation

Readings:

- CM2009: Ch.11
- Additional readings may be assigned

Lecture 11: Analysis of Panel Data

- Introduction to panel data
- Using OLS regression model on panel data
- Exploratory panel data analysis
- Unobserved effect models
- Pooled OLS models
- First-Difference models
- Distributed Lag models

Readings:

- W2016: Ch. 13
- Additional readings may be assigned

Lecture 12: Analysis of Panel Data

- Fixed Effect Model
- A Digression: differencing when there are more than 2 time periods
- Random effect model
- Fixed effect vs. random effect models

Readings:

- W2016: Ch. 14
- Additional readings may be assigned

Lecture 13: Analysis of Panel Data

- Linear mixed-effect model
- The notion of fixed and random effects in the context of linear mixed effect model
- The independence assumption

- Modeling random intercepts, slopes, and both random intercepts and slopes
- Mathematical formulation, estimation, model diagnostics, model identification, model selection, assumption testing, and statistical inference

Readings:

- JF2016: Ch. 23
- BMBW
- Additional readings may be assigned