Master of Information and Data Science
UC Berkeley

Statistical Methods for Discrete Response, Time-Series, and Panel Data (DataSci W271)

Course Designer and Developer: Dr. Jeffrey Yau (jyau@berkeley.edu)

Live Session Instructors: Jeffrey Yau (jyau@berkeley.edu)
Gerard Kelly (gkelly@berkeley.edu)

Teaching Assistant: Rahul Vaswani (rvaswani@berkeley.edu)

Office Hours:

- See the course Github Page

Last Updated: 2019 April

PLEASE READ THE SYLLABUS CAREFULLY!
Note: This syllabus is subject to change during the semester

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.

Because this course is fast-paced, mathematical notations are used throughout, and many of the concepts are quite abstract and require time to digest, students are expected to put in a substantial amount of effort to master the techniques covered in this course. It is not uncommon for students to devote 15 or more hours on average per week to this course.

Prerequisites:

1. DataSci W203 with a very solid understanding of the following materials
    - The probability and mathematical statistic concepts and techniques covered in *Probability and Statistics for Engineering and the Sciences* by Jay Devore
    - Classical linear regression modeling covered in *Introductory Econometrics* by Jeffrey Wooldridge, Chapters 1–9, Appendices A–E
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.

Students coming to the course without satisfying these prerequisites may find the course very difficult to follow and should consult MIDS program's Student Affairs Department to ensure that this is a realistic commitment.

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 in-class discussion are conducted. Attendance and participation in live session are mandatory.

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 ISVC 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-of-chapter 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 real-world 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 take considerably more time. The readings are dense and long. Depending on prior knowledge and experience, some students may have to spend significantly more time than the average amount stated above, though some students may spend less time. **If you are employed full-time and take this course together with another MIDS course, you should consult MIDS program's Student Affairs Department and ensure that this is a realistic commitment.**

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 wrong expectations (that every single proof of the theorem or derivation be provided or specific techniques be taught) and left with disappointment.

Most importantly, MIDS is a professional master's degree program, and we expect students to behave professionally. For questions regarding the course, especially those related to the materials covered in the video lectures and assigned readings, we encourage the students to use our course's slack channel. You may also post suggestions, feedback, or other topics of

discussions, but please do so using appropriate language. The use of unprofessional language in live session, e-mails, and messages on the wall will be considered misconduct and will be reported to MIDS Director of Student Affairs and MIDS program director.

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. **[HA]** Rob J Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. https://otexts.com/fpp2/
   - 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.
3. [**CM2009**] Paul S.P. Cowpertwait and Andrew V. Metcalfe. *Introductory Time Series with R*. Springer. 2009. (ISBN-10: 978-0-387-88697-8)
4. **[W2016]** Jeffrey Wooldridge. *Introductory Econometrics: A Modern Approach. 6th edition*. Cengage Learning. (ISBN-10: 130527010X)
5. **[BMBW]** Douglas Bates, Martin Machler, Benjamin Bolker, and Steve Walker. *Fitting Linear Mixed Effect Models Using lme4*
6. Additional papers, articles, and readings may be provided throughout the course.

Grade Assignment:

Total Course Score	Course Grade
[93, 100]	A
[90, 93)	A-
[85, 90)	B+
[80, 85)	B
[70, 80)	B-
[60, 70)	C
[50, 60)	D
[0, 50)	F

Total Course Score
= Total Lab Score + Total Live Session Quiz Score
+ Total Homework Score

Total Course Score Breakdown:

4 Labs	80% (20% each)
10 - 12 Homework	10%
10 – 12 Live Session Quizzes	10%

Labs

	Materials Covered	Assigned On	Due
Lab 1	Lectures 1 - 3	Monday of Week 2	4pm Pacific Time of the Monday 2 weeks later
Lab 2	Lectures 4 and 5	Monday of Week 4	4pm Pacific Time of the Monday 2 weeks later
Lab 3	Lectures 6 – 10	Monday of Week 9	4pm Pacific Time of the Monday 2 weeks later
Lab 4	Lecture 11 – 13	Monday of Week 11	4pm Pacific Time of the Monday 2 weeks later

For each of the labs, you can either work individually or in **a group of no more than three students in your same session**, though you are encouraged to work in a group. Each submission needs to include **(1) a report (in PDF format) detailing your solutions** and **(2) an RMD file (i.e. R markdown file) used to generate the solutions**. There will also be a specific page limit in the pdf report in each of the labs. Note that

- Late submission will not be accepted.
- Both of the files need to be submitted to the **Course GitHub Repo**; you can submit as many times as you want, and we will grade the last version of your submission submitted before the due date.

Each lab is due two weeks after it is assigned.

While specific instructions will be given in each lab, a few best practices in completing the labs in this course apply. **Unless specified otherwise**, a lab typically would include

- An **introduction section** that summarize the major question being asked, the methodology employed (including the final model specification), and a highlight of the results. This is true even if some of the labs are broken down by many task-oriented questions. The reason is that data science projects in real world are often tackled by breaking it down into many task-oriented questions.
- A **thorough analysis of a given dataset**, which include the examination of anomalies, missing values, potential of top and/or bottom code, etc. in each of the variables in the given dataset is provided.
- A comprehensive **Exploratory Data Analysis (EDA)** analysis, which includes both graphical and tabular analysis, as taught in this course.
    - This section is crucial in statistical modeling.
    - Output-dump (that is, graphs and tables that do not come with explanations) will result in a very low, if not zero, score.
    - Since the report has a page-limit, you will have to selectively include the visuals that are most relevant for the analysis and concise explanation of the visuals. Please do not ramble.
    - Please remember that your report will have to "walk your audience through" your analysis.
- Where applicable, a modeling section that include a detailed narrative. Make sure that your audience (in this case, the instructors, T.A.s, and your classmates) can easily follow the logic of your analysis that leads to your final model.
- Where applicable, the rationale of decisions made in your modeling, supported by sufficient empirical evidence. Use the insights generated from your EDA step to guide your modeling step, as we discussed in live sessions.
- Where applicable, all the steps used to arrive at your final model; these steps must be clearly shown and explained.
- A **conclusion** that summarize the final result with respect to the major question being asked and key takeaways from the analysis.

Homework:
Homework is graded on 0 – 3 scale. The homework file will be distributed one week before the corresponding live session. Students need to submit their solutions to the **Course GitHub Repo** before the corresponding live sessions. **Late submission will not be accepted.** Each submission needs to include (1) a report (in PDF format) detailing your solutions and (2) an RMD file (i.e. R markdown file) used to generate the solutions. You can submit as many times as you want (before the deadline), and we will grade the most updated version of your submission submitted before the due date.

- Note that there is no homework for week 1 and 14
- Each student must submit their individual homework; homework is not a group project, though you can discuss the homework with your classmates.
- The homework is graded by effort in the scale of 0 – 3:
    - A score of 3 is given if 90% or more of the questions are answered with good effort **(Note that "I gave it a try, but I got error messages so that my code didn't run" is not considered good effort.)**
    - A score of 2 is given if 70% - 90% of the questions are answered with good effort
    - A score of 1 is given if 50% - 70% of the questions are answered with good effort
    - A score of 0 is given if either no more than 50% of the questions are answered with good effort or a homework is not submitted

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. We will post our office hours on the course's Github page. Office hours are designed to discuss materials covered in the asynchronous lectures, assigned readings, and discussions in the live sessions. We encourage you to join us in the office hours to discuss course related materials. **While I am personally happy to discuss any data science related questions, the priority will be given to course related questions.** 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.

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/.

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:

Note: Instructors may provide additional materials, such as additional reading materials or new concepts (in the live sessions), during the semester.

Part 1 (Week 1 – 5): Discrete Response Models

- Bernoulli, Binomial, Multinomial, and Poisson probability distributions
- Maximum likelihood estimation
- Profile likelihood ratio test
- Inference for the probability of an event and the use of Wald, Wilson, Agresti-Coull, and Clopper-Pearson confidence intervals
- Odds, relative risks, and odd ratios
- Binary logistic regression model
- Multinomial logistic regression model
- Poisson regression model
- Hypothesis testing for regression parameters
- Log-odds of an event and its relationship to binary logistic regression models
- Probability of an event in the context of binary logistic regression models
- Variable (nonlinear) transformation and interactions
- Contingency tables and the associated inference procedures
- Test for independency
- Model specification
- Model evaluation
- Model selection

Part 2 (Week 6 – 10): Time Series Models

- Common time series patterns
- Autocorrelation and partial autocorrelation
- Notions and measures of stationarity
- Exploratory time series data analysis
- Time series regression
- Akaike's Information Criterion (and its bias corrected version) and Bayesian Information Criterion (BIC)
- Model selection based on out-of-sample forecast error
- Time series smoothing and filtering techniques
- Stationary and non-stationary time series processes
- Stationary Autoregressive (AR), Moving Average (MA), and Mixed Autoregressive Moving Average (ARMA) processes
- ARIMA model
- Seasonal ARIMA model
- Estimation, diagnostic checking of model residuals, assumption testing, statistical inference, and forecasting
- Regression with autocorrelated errors
- Autoregressive Integrated Moving Average (ARIMA) Model
- Unit roots, Dickey-Fuller (ADF) test, and Phillips-Perron tests
- Spurious regression and Co-integration
- Vector Autoregressive (VAR) Models

Part 3 (Week 11 – 13): Statistical Models for Panel (or Longitudinal) Data

- Exploratory panel data analysis
- Pooled OLS regression model
- First-differenced regression model
- Distributed lag model
- Fixed-effect regression model
- Random-effect regression model
- Linear mixed-effect model

Detailed Course Outline:

Lecture 1: Discrete Response Models

- 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 Models

- 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 Models

- 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 Models

- 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 Models

- 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, 5.1 - 5.4
    - Skim sections 5.2.3, 5.3
- 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
- Additional readings may be assigned

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

Reference Readings:

- CM2009: Ch. 3 and 5
- 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: 3, 4.5, 6
- HA: the sections that cover the same materials this week
- 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 and 7
- HA: the sections that cover the same materials this week
- 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
- HA: the sections that cover the same materials this week
- 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:

- BMBW
- Additional readings may be assigned