Master of Information and Data Science
UC Berkeley

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

Syllabus

Last Updated: 12/5/2016

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

Course Description:
This course teaches data scientists a number of statistical models at the data-science industry-professional level. It teaches 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. This is a fast-paced course that studies statistical models for the analysis of cross-sectional data with unordered and ordered categorical variables, count response variables, univariate time-series data, multivariate time-series data, and panel data. Emphasized throughout are exploratory data analysis (EDA), how to use the insights gained from EDA in subsequent statistical modeling applied to solving data science problems that are often given as (vague) business or policy problems, mathematical formulation of statistical models, assumptions underlying these models, the consequence when one or more of these assumptions are violated, the potential remedies when assumptions are violated, model selection, model diagnostic, model assumption testing, and model evaluation. The course goes well beyond the mechanical implementation of statistical methods using 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 and mathematical statistics 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 a substantial amount of effort to master the techniques covered in this course.

Prerequisites:
1. DataSci W203 with at least a grade of B+
    - 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.
    - Differential calculus, integral calculus, matrix notations, matrix differentiation, and integration are used extensively throughout the course.

Students coming to the course without satisfying these prerequisites will find the course very difficult to follow and must seek permission to enroll from Professor Jeffrey Yau.

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. Participation in live session is mandatory. Students coming to live sessions unprepared and not actively participating in discussions will have their participation points deducted. Students should have a place with good Internet connection to attend the live session. If you mute your video during the live session, the professor will call you out. Students are expected to actively participate in the live sessions 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 to 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. The textbooks go into a lot more details than the asynchronous lectures and provide many more examples that are not possible to cover in an 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.

This is a fast-paced course, and the mathematical structure and assumptions of the statistical models taught are covered in-depth. 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, students should expect to spend on average at least 15 to 20 hours per week in this course outside of watching the asynchronous lectures and attending the live sessions. There are weeks 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 expected amount stated above, though some students may spend less time. If you are employed full-time, you are highly discouraged from taking this course together with any other courses. If you are employed full-time and take this course together with another MIDS nonelective course, you will have to seek permission from Professor Jeffrey Yau. If you are employed full-time, you are not permitted to take this course with another MIDS elective course.

Note that this is not a graduate-level mathematical statistics course. This is an advanced statistics course for aspiring data scientists or data scientists who want to learn the foundational statistical techniques to model categorical, time-series, and panel data. 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 aspiring Ph.D. statisticians or econometricians. Do not take this course if you expect every single mathematical proof related to the models studied in this course be provided.

Most importantly, 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 the ISVC wall. 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.

Communications between Instructors and Students:
For questions regarding the course, please use the ISVC wall so that other students can see both the questions and answers provided by instructors, other students, and, in some rare occasions, the course developers. While our teaching team meets on a weekly basis and has frequent communication, your live session instructors should be the main point of contact regarding the materials of this course.

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. **[SS2016]** Robert H. Shumway and David S. Stoffer. *Time Series Analysis and Applications. EZ Edition with R Examples.* Version 2016.4. www.stat.pitt.edu/stoffer/tsa4/tsaEZ.pdf
3. **[HA]** Rob J Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice.* https://www.otexts.org/fpp
4. [**CM2009**] Paul S.P. Cowpertwait and Andrew V. Metcalfe. *Introductory Time Series with R*. Springer. 2009. (ISBN-10: 978-0-387-88697-8)
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. Additional papers, articles, and readings will be provided throughout the course.

Note: this is a new version of w271 class. This syllabus is subject to change, and changes will be announced on the ISVC course wall.

Grade Assignment:

[94, 100]	A
[90, 94)	A-
[85, 90)	B+
[80, 85)	B
[70, 80)	B-
[60, 70)	C
[50, 60)	D
[0, 50)	F

Grading:

Class Participation (including live session attendance)	10%
Labs 1 and 2	15% each
Labs 3 and 4	30% each

Lab Due Dates:

- Lab 1, which covers Lectures 1–3, is due on the Sunday of Week 5
- Lab 2, which covers Lectures 4 and 5, is due on the Sunday of Week 7
- Lab 3, which covers Lectures 6–10, is due on the Sunday of Week 11
- Lab 4, which covers Lectures 11–13, is due on the Sunday of Week 14

There are four labs. You can either work individually or in a group of no more than four students, though you are strongly encouraged to work in a group. Each submission needs to include (1) a report (in PDF format) detailing your solutions and (2) a R script or RMD file that you use to generate the solutions.

- Failing to submit one of these files will receive an automatic 50% reduction in the grade.
- Failing to follow the instructions, including the file naming convention, specified in the labs will receive 20% reduction in the grade.
- Late submission will not be accepted.
- Files will be submitted to the corresponding links on the ISVC platform.

Participation will be graded based on attendance and contribution to the discussions in the live sessions. Note that we have a "no-muting-video" policy. Make sure you have good internet connection when attending the live sessions and you are not allowed to mute your video.

Course Outline:

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 logistical 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 selection and evaluation

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)
- 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 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

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
- More readings to 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
- More readings to 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 the following sections 5.2.3, 5.3
- More readings to 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, 2.1.1, 2.2.4, 2.2.5, 2.3, 4.2
    - Skip Ch. 1.5.4, 1.5.5
- SS2016: Ch.1
- More readings to 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

Readings:

- CM2009: (Optional) Ch. 3.4 "Exponential Smoothing"
- CM2009: Ch. 5.1 – 5.3
- SS2016: Ch.2
- HA: (Optional) Ch.7 "Exponential Smoothing"
- More readings to 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: 3.1, 3.2, 4.5, 6.1 – 6.4
- SS2016: Ch. 3.1 – 3.6
- HA: Ch. 8.2, 8.3, 8.4
- More readings to 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 – 4.7, 6, 7.1 – 7.3
- SS2016: 3.7 – 3.10, review 3.1 – 3.6
- HA: Ch. 8.5 – 8.9
- More readings to 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
- SS2016: Ch. 5.3, review Ch. 2.1
- HA: Ch. 9.2
- More readings to 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 (skip 13.4)
- More readings to 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. 14.1 – 14.2
- More readings to 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 / forecasting, backtesting

Readings:

- BMBW
- More readings to be assigned