Applied Regression and Time Series Analysis Syllabus

2016 Summer

Course Developers:	Jeffrey Yau jyau@ischool.berkeley.edu	Paul Laskowski paul@ischool.berkeley.edu
Instructors:	Samuel Frame drsamueljframe@gmail.com	Devesh Tiwari devesh.tiwari@gmail.com

Office Hours: To-be-posted by the Instructors on ISVC

Course Description:
Classical linear regression and time series models are workhorses of modern statistics, with applications in nearly all areas of data science. This course takes a more advanced look at classical linear (regression) models and introduces the fundamental techniques of time series modeling, focusing on the class of univariate linear time series models [*]. The first half of course covers the classical linear model and the second half covers time series models. While the course covers implementation of the models using *R*, this course is not just about the mechanical implementation using *R*. Mathematical formulation of statistical models, assumptions underlying these models, the consequence when one or more of these assumptions are violated, and the potential remedies when assumptions are violated are emphasized throughout.

Major topics include classical linear regression modeling, casual inference, identification strategies, and a class of univariate time series models that is very popular among industry professionals. Throughout the course, we emphasize formulating, choosing, applying, implementing, evaluating, and reformulating statistical techniques to capture key patterns exhibited in data. All of the techniques introduced in this course come with examples using simulated or real-world data, and some examples also come with *R* codes. Students who successfully complete this course will be able to decide on fwhat techniques are appropriate for a given question, and to make trade-offs between model complexity, ease of interpreting results, and implementation timing in real world applications. As concepts in probability theory and mathematical statistics are used extensively, students should feel comfortable with the definition, manipulation, and application of these concepts in mathematical notations.

Prerequisites:

1. DataSci W203
2. Hands-on experience in R
3. Working knowledge of calculus. Though not required (as this course does not present statistical models using matrix algebra), a good understanding of linear algebra and linear difference equations would be helpful.

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 before attending the live sessions, where group exercises are assigned and in-class discussion are conducted. Students are expected to actively participate in the live sessions and contribute to the discussions. Students should also come to

[*] Models for longitudinal data, which in data science (unfortunately) is often referred to as *time series data*, are not covered in this course.

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 in advance. It is important to note that live sessions are not lectures, though the instructors sometimes may spent up to 15 minutes to review some key concepts covered in the asynchronous lectures or the readings. It is also important to know that the asynchronous video lectures and the assigned textbook readings are not substitute of each other. The textbooks go into a lot more details and provide many more examples that are not possible to covered 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.

Although it is not a fast-pace course, the mathematical structure and underlying statistical assumptions of both the classical linear model and the ARIMA model are covered in-depth. While we cover the mechanic of implementing these models in *R*, the course, which focuses on building statistical models that can be applied to real-world data science problems, is not just about the mechanics. In fact, many of the *R* libraries introduced in this course have many 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.

When making your schedule for this course, note that there are weeks, notably those in which a lab is assigned, may take considerably more time. Also, depending on prior knowledge and experience, some students may spend a lot more and some may spend a lot less than expected.

This is not a graduate level mathematical statistics course. This is an *applied regression and introduction to time series analysis* course for aspiring data scientists. This course emphasizes on data science applications of statistical techniques, and 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 or even presentation using matrix notations, 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 are very comfortable with linear algebra and concepts in stochastic convergence. Therefore, students should expect that this course is designed for aspiring data scientists and not for aspiring Ph.D. statisticians or econometricians.

Most importantly, we expect the 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 with appropriate language.

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 occasions, the course developers. While our entire teaching team meet on a weekly basis and have very frequent communication, your live session instructors should be the main point of contact regarding the materials of this course.

Required Textbooks:

1. **W2012** Jeffrey Wooldridge. *Introductory Econometrics: A Modern Approach.* 5^{th} *ed.* Cengage Learning. 2012

2. **CM2009** Paul S.P. Cowpertwait and Andrew V. Metcalfe. *Introductory Time Series with R*. Springer. 2009.

3. **CR** Course Reader, which include 2 chapters from **(D2008)** Zoltan Dienes, *Understanding Psychology as a Science*, Palgrave, 2008.

Course Outline:

1. A Review of Key Concepts in Probability and Statistics (2 lectures) The course begins with a compact review of probability theory, followed by a discussion of the foundations of Frequentist and Bayesian statistics.

    - Probability density and cumulative functions, marginal, joint, and conditional probability
    - Notion of random variables
    - Unconditional and conditional expectation
    - Variance and covariance
    - Statistical estimation and desirable properties of estimators
    - Frequentist approach to statistical inference
    - Bayesian approach to statistical inference

2. The Classical Linear Model (3 lectures) We will spend three weeks on the fundamentals of linear regression (describing the topics with much more rigor and detail than those covered in **w203**) and show how an assumption of linearity allows for a great variety of model specifications.

    - Classical Linear Model Assumptions
    - OLS estimation mechanics
    - Statistical Inference
    - The use of variable transformation, polynomials, indicator variables, and interaction terms
    - Regression Diagnostics and formal statistical assumption testing

3. Causal Inference and Identification (2 Lectures) We discuss the application of linear models to causal inference. We use canonical studies as examples of how researchers leverage natural experiments, aim for an intuitive understanding of identification strategies, and apply these strategies in real world examples.

    - Causality, the notion of exogeneity, and omitted variable bias
    - True experiments and Natural experiments
    - Instrumental Variable Method
    - An Overview of Regression Discontinuity

4. Time series Statistical Models (7 Lectures) The second half of the course covers some of the most fundamental techniques in statistical time series modeling.

    - Exploring and visualizing Time Series data
    - Time Series Regression and the No-serial Correlation Assumption under CLM
    - Time Series Smoothing and Filtering Techniques
    - Stationary and Non-stationary Time Series Processes
    - Stationary Autoregressive (AR), Moving Average (MA), and Autoregressive Moving Average (ARMA) processes
    - Estimation, Diagnostic checking of model residuals, Assumption Testing, Statistical Inference, and Forecasting
    - Autoregressive Integrated Moving Average (ARIMA) Model, Unit roots, Dickey-Fuller (ADF) and Phillips-Perron tests
    - Regression with Time Series Data, Spurious regression, and Co-integration

- An Introduction to Vector Autoregressive (VAR) Models
- An Introduction to Generalized Autoregressive Conditional Heteroskedastic (GARCH) model

Grading Rubic[†]**:**

Score Range	Grade
[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:

1. Quizzes and Class Participation - 10%
2. 3 Group Labs - 60% (lab 1 - 5%, lab 2 - 25%, lab 3 - 30%)
3. 1 Midterm Exam (Online) - 15%
4. 1 Final Exam (Online) - 15%
5. 8 Exercises (Optional but highly recommend; they are not required to be turned in)

Tentative Schedule:

Week	Quizzes, Labs, Exams	Materials Covered	Assign	Due
1	Lab 1	week 1 and 2	week 1	week 3
2				
3				
4				
5				
6				
7	Lab2	week 2 - 7	week 7	week 9
8	**Mid-term Exam (in-class)**	week 1 - 7	week 8	in class
9				
10				
11				
12				
13	Lab 3	week 8 - 13	week 12	week 14
15	**Final Exam (in-class)**	week 1 - 14	week 14	in class

Quizzes:
There are **five quizzes**. Occasionally, quizzes will be given at the beginning of a live session. The purpose of the quizzes is to test your understand of the asynchronous materials covered in specific units. Each quiz is short and is given $5 - 10$ minutes to complete.

[†] *A*+ is given at the discretion of the instructors and course developers.

Midterm and Final Exams:
There are **two exams**. They are instrumented online during the live sessions on week 7 and 14 and must be done within the time window given. Each exam is given **40 minutes to complete**.

Labs:
There are **three labs**, and these **must be done in groups**. Please let your instructor know at least a week in advance the members of your group and ask them for help if you have difficulty finding a group. Each group needs to submit **1** report (in PDF format) detailing your solutions and **2** R-script(s), juypter notebook, or Rmd file that you use to generate the solutions. Failing to submitting one of these files will receive an automatic 50% reduction in grades. Since the first two weeks focus on probability and mathematical statistics, Lab 1 requires paper-and-pencil calculations on the materials from these weeks. Lab 2 and 3 have a strong focus on empirical applications. The labs are given about **two weeks to complete**.

Optional Exercises:
Some weeks include assignments that are optional. They serve two purposes: (1) Ensure that you have learned the asynchronous material and can apply your knowledge effectively. (2) Extend the asynchronous material with new concepts, theories, and advanced techniques. All the assignments have a strong focus on empirical applications and require the use of R. Most of the R functions and packages that are needed in the assignments will first be taught in either the asynchronous R sessions, and/or additional R scripts. We encourage you to work on the exercises, although you do not have to turn them in.

Detailed Course Schedule:

Lecture 1: Probability Theory

1. Basic concepts in probability theory
2. Probability density, mass, and cumulative functions
3. Marginal, joint, and conditional probability
4. Random variables
5. Expectation and conditional expectation
6. Variance and covariance
7. Parameters and estimators
8. Bias, consistency, asymptotic normality

Readings:

- **W2012**: Appendix B
- **W2012**: Appendix C1-C4

Lecture 2: Foundations of Mathematical Statistics

1. The Neyman-Pearson approach
2. Multiple comparisons
3. Planned vs. posthoc tests
4. Stopping rules
5. Some basic concepts in Bayesian Statistics
6. Prior and posterior belief distributions
7. The likelihood principle

Readings:

D2008: Ch.3

Lecture 3: Ordinary Least Squares Estimation

1. The linear population model
2. Bivariate OLS estimation
3. Multivariate OLS estimation
4. The Gauss-Markov Theorem
5. Diagnostics and responding to assumption violations

Readings:

- **W2012**: Ch.2, pp.22 - 45
- **W2012**: Ch.3, pp.68 - 89, 95 - 104

Lecture 4: Ordinary Least Squares Inference

1. Classical linear model assumptions
2. Hypothesis testing
3. Confidence intervals

4. Responding to violations of normality
5. OLS asymptotic
6. Joint significance and model significance
7. Diagnostics and responding to assumption violations

Readings:

- **W2012**: Ch.4
- **W2012**: Ch.5

Lecture 5: Linear Model Specification 1

1. Logarithmic transformations
2. Quadratics and higher-order polynomials
3. Indicator Variables
4. Interaction terms and their interpretation
5. Diagnostics and responding to assumption violations

Readings:

- **W2012**: Ch.6
- **W2012**: Ch.7

Lecture 6: Linear Model Specification 2

1. Omitted variable bias
2. Endogeneity
3. Causality
4. Proxy variables
5. Lagged independent variables
6. True experiments
7. Natural Experiments

Readings:

- **W2012**: Ch.3, pp. 89 - 95, Appendix 3A.4
- **AP2009**: Ch.2

Lecture 7: Instrumental Variables

1. The instrumental variables method
2. Wald estimator
3. Local average treatment effect
4. Overview of regression discontinuity

Readings:

- **AP2009**: Ch.1
- **W2012**: Ch.15

Lecture 8: Introduction to Time Series Analysis

1. Real world examples from different fields, including public sector, meteorology, business, and higher education, are used to illustrate the importance of time series analysis
2. Basic terminology and fundamental concepts of time series analysis
3. Exploratory time series data analysis
4. A few elementary time series models
5. Time series simulation

Required Readings:

- **CM2009**: Ch.1, Ch.4.2.1 - 4.2.3, Ch.4.3.1- 4.3.2
- **W2012**: Ch10.1, 12.1-2

Lecture 9: Measures of Dependency, Notion of Stationarity, and Time Series Smoothing Techniques

1. Notion and Measure of Dependency of a Time Series
2. Notion of Stationarity
3. Stationary and Non-Stationary Time Series Processes
4. Measure and Estimation of Dependency
5. Basic properties and Simulation of the following models:
   (a) White Noise
   (b) Polynomial Trend
   (c) Moving Average
   (d) Autoregressive Models
   (e) Random Walk
6. Common Time Series Smoothing Techniques
   (a) Moving Average Smoothing
   (b) Exponential Smoothing
   (c) Polynomial and Periodic Regression Smoothing
   (d) Spline Smoothing
   (e) Kernel Smoothing

Required Readings:

- **CM2009**: Ch.2, Ch.4.1-4.4
- **W2012**: Ch.10.2 (but skip *Finite Distributed Lag Model*, 10.3, 10.5, 11.1)

Lecture 10: Modeling Stationary Time Series using Autoregressive (AR) Models

1. Approach to learn time series modeling in this course
2. Mathematical formulation and derivation of the properties of AR Models
3. Simulating time series using AR models
4. Exploratory Time Series Data Analysis with the use of ACF and PACF
5. Identification of the order of dependency
6. Model Estimation
7. Perform model diagnostics using estimated residuals and various graphical and statistical techniques
8. Evaluating Model Performance using both in- and out-of-sample fit

9. Cautions of using certain R functions to estimate AR models

Required Readings:

- **CM2009**: Ch.4.5-4.6, Ch.6.1-6.4
- **W2012**: Ch. 11.1, 11.2, 11.3 (but skip *"Transformations on Highly Persistent Time Series" and "Deciding Whether a Time Series is I(1)"*)

Lecture 11: Modeling Stationary Time Series using Moving Average (MA) Models and Mixed Autoregressive and Moving-Average (ARMA) Models

1. Mathematical formulation and derivation of the properties of MA and ARMA Models
2. Formulate MA model using lag operators and study its properties in the lag operator form
3. Conduct simulations (using R) and examine the empirical features of MA and ARMA Models
4. Exploratory Time Series Data Analysis with the use of ACF and PACF
5. Identify the order of dependency and building MA and ARMA models
6. Model Estimation
7. Diagnostic Checking and Testing Model Assumptions
8. Evaluating Model Performance
9. Forecasting using MA and ARMA models
10. Evaluating forecasts
11. Discuss the issues when selecting the “best” model for the question under consideration and the purpose for which the model is built

Required Readings:

- **CM2009** Ch.6.5 - 6.6

Lecture 12: Modeling Non-Stationary Time Series using Autoregressive Integrated Moving Average (ARIMA) Models

1. Transformation of a non-stationary process to a stationary process
2. Using Integrated models for non-stationary series
3. Key characteristics of random walk processes
4. Unit Roots Stationarity and Testing for Unit Roots
5. Steps to build an ARIMA Model: The Box-Jenkins Approach
6. Testing model assumptions
7. Evaluating model performance
8. Forecasting using an ARIMA Model
9. Evaluating forecasts
10. Seasonal ARIMA Models

Required Readings:

- **CM2009** Ch.7.1 - 7.3

Lecture 13: Regression with Time Series Data and An Introduction to Vector Autoregressive Moving Average Models

1. Identify time series with stochastic trends

2. Spurious correlation and spurious regressions
3. Regression with time series data
4. Notion of unit roots
5. Detecting unit roots using Augmented Dickey-Fuller Test
6. Notion of cross-correlation, autocorrelation, and lead-lag relationship in multivariate time series
7. Vector Autoregressive (VAR) Models: An Introduction
8. Regression Models with time series errors

Required Readings:

- **CM2009** Ch.11
- **W2012**: Ch.12.3-12.6, 18.2-3, 18.4 (but skip *"Error Correction Model"*)

Lecture 14: An Introduction to Generalized Autoregressive Conditional Heteroskedastic (GARCH) models

1. Relaxing the constant variance assumption: Understand the motivation for modeling variance dynamics
2. The ARCH process in mathematical form
3. The GARCH process in mathematical form
4. ARCH and GARCH Model Estimation
5. Testing model assumptions
6. Evaluating model performance (both in- and out-of-sample)
7. Forecasting using an ARIMA/GARCH Model
8. Evaluating forecasts
9. Drawbacks of the base GARCH Model

Readings:

- **CM2009** Ch.7.4