STAT 361/661a: Data Analysis

Lisha Chen (lisha.chen@yale.edu)

Fall 2013

Practical Information

- Lectures: Monday and Wednesday 2:30-3:45; Location: DL 220.
- TF: Saier (Vivien) Ye (saier.ye@yale.edu)
- Office hours: TBA
- Please join in the class website on classesv2! Class Materials are available at under "resources".

Prerequisites and Objectives

This course is intended for advanced undergraduate students, first year graduate students in statistics, and graduate students from other disciplines who are interested in applied statistical methods. It is helpful to have some background in probability and statistics, such as a course on the level of STAT 242 or STAT 238, and some experience in programming. FAQ section toward the end of this syllabus answers some common questions about the prerequisites.

Typical steps for data analysis:

1. collecting and organizing the relevant dataset; asking questions about the data context.
2. exploratory data analysis and data visualization.
3. building statistical models.
4. summarizing and interpreting the results.

In the lectures, we will mostly focus Steps 2-4, as we will mostly use classical and clean datasets for class demonstration. We will cover a wide range of statistical methods with both simulation and real data examples for illustration and use R software for implementation. Step 1 is a very important step, which should be a substantial component in the final projects. I hope that through this course you will understand the mechanism of the statistical methods and practical considerations of data analysis, and develop a set of data visualization, statistical modeling and computing skills that enable you to adapt to the inevitable surprises of data.

Topics

- Exploratory data analysis and data visualization
- Permutation test and bootstrap
- Linear regression
- Logistic regression and generalized linear models

- Generalized Additive Models
- Tree based methods
- Maximum Likelihood and EM algorithm
- Principal components analysis and clustering analysis
- Time Series Analysis

Grading

Grades will be based on homeworks (65%), a final project (30%), and class participation (5%).

- Homework will be assigned once every other week but will require a regular amount of work. You will benefit from starting to work on the homework during the first week it is assigned.
- For the final project, you have the opportunity to choose a dataset of your own interest to work on. You can work independently or pair up with one other student in class. Keep an eye on real data opportunity from the beginning of the semester. You will be asked to do a midterm presentation and a final presentation on your project.
- You are also encouraged to contribute to the class by actively participating in the class and giving feedback on class materials.

Auditors: To audit this course, you are required to attend the class regularly and participate in class discussion. Please talk me at the beginning of the semester so that I am aware of your status.

Computing

Please bring your laptop to class. We will learn and use the R statistical programming environment. R is an Open Source implementation of S language and it is free downloadable at http://www.r-project.org/.
Computing facilities are available at the Statlab (219 Prospect St. CSSSI), and R is installed on Yale's cluster PCs. CSSSI also offers R workshops during the semester and consultation with R and data analysis.
(http://statlab.stat.yale.edu/people/showConsultants.jsp)

References (and recommendations)

On Statistical Methods (with examples illustrated by R/S) :

- Julian J. Faraway, "Linear Models with R". A clear and concise book on linear models with many data examples for illustration. Some class materials on linear regression are adopted from this book. (Highly recommended)
- Maindonald and Braun, "Data Analysis and Graphics Using R – an Example Based Approach". An exposition of statistical methodology that focuses on ideas and concepts based on R. Recommend the book to students who have already had a regression or a linear models course.

- Dalgaard, “Introductory Statistics with R”. Elementary-level introduction to data analysis and R. Suitable for students with relatively less background in statistics.
- Venables and Ripley, “Modern Applied Statistics with S”. A book on a broad array of applied statistical topics based on S-Plus. The book emphasizes more on the language rather than methodology. It can be used as a reference of R/S-Plus for advanced users.
- Chambers and Hastie, “Statistical Models in S”. On both the statistical modeling and the theory underlying the S/R functions. Also called “the white book”. Recommend for in-depth understanding on statistical modeling and S/R langue.

On Theoretical Foundations: Recommended textbooks for students who have not taken 242 or 238 and look for a better understanding of the theoretical foundations.

- John Rice, “Mathematical Statistics and Data Analysis”. The textbook for Stat 242/542 (Theory of Statistics). Basic and fundamental theory of probability and statistics with good examples.

- Jay L. Devore and Kenneth N. Berk, “Modern Mathematical Statistics with Applications”. Provide a balance between theoretical foundations and statistical practice. E-book available through Yale Library. http://www.springerlink.com/content/978-1-4614-0390-6/#section=999513&page=5&locus=90

- Weisberg, “Applied Linear Regression”. A classical textbook on linear regression

On programming in R:

- A easily accessible R reference: “Quick-R” at http://www.statmethods.net/ . Provide concise illustration of the mostly commonly used R functions for different types of analysis.
- An Introduction to R, **FREE** downloadable at http://www.r-project.org/ (under documentation -> manuals).

Shopping this class and homework

You are welcome to shop the class. However, all homework assignments (including the first one) are required of everyone, submitted on time. This will not be an easy course to slip into after the first class.

Frequently Asked Questions

Following are my responses to some commonly asked questions about this course. You are always welcome to talk to me in person about your background and concerns.

1. Q: Can I take this course without taking stat 238 or stat 242?
   A: Stat361 emphasizes not only on how the statistics analysis should be done but also why it needs to be in done a particular way. We would like to obtain conceptual understanding of the data analysis methods to apply them appropriately. Therefore some previous exposure to fundamental concepts and

theory of probability and statistical inference will be helpful. For example, you should be fairly familiar with the following concepts

probability distributions (such as normal distribution, Bernoulli distribution), expectation, variance, standard deviation, confidence intervals, likelihood, hypothesis testing, test statistics, null distributions, p-values …

Ideally stat242 or stat238 is taken before stat361 so that you can get most out of the course. However stat242 or stat238 is not a hard requirement. Other statistics course may also serve the purpose. You are encouraged to shop the course and see how you feel.

2. Q: Can I take this course without any experience in R?
   A: Yes, prior knowledge in R is not required. But it would be helpful that you have had some programming experience in other languages, such as C/C++, matlab, java, Perl, SAS ect, which will help you pick up R pretty quickly. Part of the course requires understanding of basic programming concepts, such as conditional statements and loops ect.

3. Q: I have taken introductory data analysis stat 230. Is there too much overlap between stat230 and stat361?
   A: There is certain amount of overlap. For example, linear regression is covered in both courses. Some advanced topics like tree-based models, and generalized additive models might not be covered in the introductory data analysis. However the main difference is that stat 361 is focused more on the methodological aspect of the data analysis tools and demonstrate the methods using real world data, whereas stat230 emphasizes more on practice.

Acknowledgement

Many thanks go to my Ph.D. advisor Dr Andreas Buja for generously sharing his course materials.