Introduction to Data Analysis with R

Mon, Wed, 3-4:14, ECCR 143 Spring 2016.

Douglas Nychka
Affiliate Faculty, Applied Mathematics
Scientific Staff, National Center for Atmospheric Research

Introduction

Data science is a new field that combines mathematics, computer science and statistics with the goal of answering questions and discovering new information from data. Some examples are: How should I buy a used car on cars.com? Where will the next flu outbreak occur? How is the climate changing in Colorado? What is the electrical power consumption of a supercomputer? These kind of practical questions require skills in wrangling data sets into forms for analysis, using graphics to explore relationships among variables, writing programs to do statistical analysis, and communicating the results in nontechnical language.

The goal of this course is to expose students to data analysis and discovery using data science. In the process students will learn how to write programs in the R language and generate figures and reports. R is a community-based data analysis environment that is free and a standard in data science and statistics. It is a flexible language similar to Matlab and python, is fun to use, and is good for learning how to program. The course will also introduce students to some modern data analysis tools including regression and smoothing, multivariate analysis, clustering, spatial prediction and image analysis. Where it is appropriate students can adopt data sets and project that follow their own interests and majors.

Some of the statistical and mathematical background for the analysis techniques will be given but the emphasis will be on solving real data problems and learning how to work with these methods in R. We will take the approach that many sophisticated and advanced methods can be appreciated and used within the context of particular data sets if students have a clear ideas of the analysis goals and an understanding of how the data is collected or generated. This kind of understanding is a practical complement to the more mathematical development that would occur in other statistics or computer science courses.

Course Goals

1. Become skillful writing programs and generating reports in R and RStudio.
2. Learn how to explore and check data sets using graphics and other statistics.
3. Develop skills in manipulating and transforming a data set into more useful formats.
4. Develop skills in visualizing and summarizing univariate and multivariate data.
5. Learn how to match a statistical method to a particular type of data set or question.
6. Develop skills for communicating the results of a data analysis.

Course Outline

For most weeks the first lecture will present some new topics with subsequent lectures focusing on hands-on exercises in class. The grading will be based on weekly homework and several (short) take home quizs, and project to be presented in class. Data analysis and programming are skills that are developed by doing. The strategy is to have many short and regular points of evaluation instead of a few big tests. The take home quizzes will typically feature a few practical questions for a specific data set and will allow students to use their own PC, use the web, and also give more time to think about solutions. For many parts of the course, students will be encouraged to work together in small groups to develop collaborative skills and in particular the final project can be done as a team.

Weeks 1 - 3	Data programming in R
	Quiz 1: Creating simple functions for data analysis.
Weeks 4-6	Manipulating and Plotting Data
	Quiz 2: Formating and interpreting a data set.
Weeks 7- 9	Regression analysis and curve fitting
	Quiz 3: Regression analysis
Weeks 10-12	Clustering
	Quiz 4: Multivariate analysis
Weeks 13-15	Spatial and image data
	Project: Communicating a data analysis