MATH 4470/5470 EXPLORATORY DATA ANALYSIS
ON-LINE
FALL 2014

Text

There is no text for this course. All of the learning material is available on the lecture notes. This book is based on John Tukey's text *Exploratory Data Analysis*, although you do not need to read this book.

Almanac

Although there is no required textbook, a copy of any current world almanac may be helpful to purchase or use in the library. This is relatively inexpensive and it provides a convenient source for data for this course. For many of the homework assignments, you will be asked to work on your own data. If you dont want to purchase an almanac, then you will be able to get data from the Internet.

Goals of the class

This class is designed to give the student a basic understanding of the principles behind modern data analysis. One can think of data analysis as numerical detective work. Given a batch or several batches of data, we wish to discover the basic pattern or structure that is present. Also, by looking at residuals, we look for interesting structure that can only be viewed after one has removed the basic patterns from the data. The student will learn some novel techniques for exploring data. More importantly, he or she will learn a basic philosophy of data analysis that will guide any type of data exploration in the future.

Outline of Topics

1. Introduction to EDA
2. Working with a single batch
3. Comparing batches
4. Transformations
5. Reexpressing for symmetry
6. Resistant line
7. Straightening plots
8. Smoothing sequences
9. Two-way tables median polish
10. Binned data – rootograms
11. Fraction data

Homework

A typical homework assignment will consist of several data analyses following the techniques described in the lecture notes. **All homework assignments are to be turned in electronically by html files created by R Markdown available through RStudio.**

Project

One component of this class is a data analysis project. You choose an interesting dataset and do an extensive analysis using methods from the course. You turn in a 5-10 page report that describes the dataset and what you learned about the structure of the data in your analysis.

Grading

Each of the 8 main homework assignments is worth 20 points, and each of the 6 activities and Homework 0 is worth 10 points. The project is worth 50 points. Letter grades of A, B, C, D correspond to 90%, 80%, 70%, and 60% of the total points. Unexcused assignments turned after the due dates will be penalized. There will be a penalty of 4 points for each day late assignments that are more than two days late will not be accepted.

Computing using R

We will be using the statistical system R and the interface RStudio for performing the EDA methods that well use in this course. R is a free package that will run on all platforms (Windows, Macintosh, or Unix.) I have written a R package LearnEDA (version 1.4) that contains functions and datafiles for this class. This package is only available by downloading the compressed file from a link from the home page of the instructor. I will be providing basic instruction on R in the first weeks of the class. The activities are designed to illustrate graphically and dynamically (using R) many of the concepts in the course.