

MATH 447/547 EXPLORATORY DATA ANALYSIS

ONLINE COURSE SYLLABUS

COURSE DESCRIPTION

Introduction to modern techniques in data analysis, including stem-and-leaves, box plots, resistant lines, smoothing and median polish.

SCHEDULE

Fall of odd years. 3 credit hours.

PREREQUISITE

C or better in MATH 247, MATH 341, or MATH 441 or consent of instructor.

COURSE PURPOSE

This class is designed to give the student a basic understanding of the basic 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.

The course discusses the display and summary of one batch, the effective comparison of batches, relationships between two measurement variables, smoothing time series data, additive fits to two-way tables, and working with binned data.

Although we discuss a variety of methods useful for exploring data, they all share different characteristics that underlie the philosophy of Exploratory Data Analysis (EDA).

Revelation: In EDA, there is an emphasis on using graphs to find patterns or displaying fits. Effective displays of data can communicate information in way that is not possible by the use of numbers.

Resistance: In EDA, we wish to describe the general pattern in the majority of the data. In this detective work, we don't want our search to be unusually

affected by a few unusual observations. So it is important that our exploratory method is resistant or insensitive to outliers.

Reexpression: We will see that the natural scale that the data is presented is not always the best scale for displaying or summarizing the data. In many situations, we wish to reexpress the data to a new scale by taking a square root or a logarithm. In this class, we will talk about a useful class of reexpressions, called the power family, and give guidance on the “best” choice of power reexpression to simplify the data analysis.

Residual: In a typical data analysis, we will find a general pattern, which we call the FIT. The description of the FIT may be very informative. But in EDA we wish to look for deviations in the data from the general pattern in the FIT. We look at the residuals, which are defined as the difference between the data and the FIT.

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

REQUIRED RESOURCES

Text

There is no text for this course. (The material is learned using the on-line lecture notes.) A copy of Tukey’s text Exploratory Data Analysis will be kept on two-hour reserve in the Math-Science library. This book is similar in spirit to this course and contains a discussion of the exploratory methods that we will be using.

Almanac

Although there is no required textbook, I highly recommend that you purchase a copy of any current world almanac. 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 don’t want to

purchase an almanac, then you will be able to get data from the Internet or the libraries on campus.

Internet Access

All of the content of the course can be found on the BGSU Blackboard site. This includes all of the lecture notes, tutorial material on the R and Fathom software, and all homework assignments and Fathom lab assignments.

COURSE DELIVERY

Computing using R and Fathom

We will be using the statistical software R for performing the EDA methods that we'll use in this course. R is free software. You may download it on your own computer. Also, several labs will have R available on campus. In addition, there will be a number of labs that use the Fathom software – these labs are designed to illustrate graphically and dynamically many of the concepts in the course. Fathom is also available on-campus. You can purchase a student copy of Fathom from Key College (www.keycollege.com) for your home computer.

Organization of Blackboard material:

- * EDA Lectures
Here you find all of the lecture notes for the course.
- * R STUFF
Here I will put tutorials and demonstrations of data analysis commands using the R software.
- * FATHOM
Here I will put all of the Fathom labs.
- * DATA
Here I will place all of the datasets that we will use in this class (lectures and homework).
- * HOMEWORK
All of the homework assignments are located here
- * CLASS SCHEDULE
This tells you the pace at which you should be working through the material and the homework assignment dates.

Distance Learning and Communication

This is a non-traditional class – there will be no regularly scheduled classes on-campus and the idea is to learn the material by reading the lecture notes and doing homework. The pace at which you work is shown in the timetable in the Class Schedule. The instructor will use various methods to communicate with

the students regarding any problems that arise. There will likely be weekly help sessions held on campus to go over software and homework problems. The best way of communicating with the instructor is through email and the instructor will respond to any email within 24 hours

COURSE GOALS

Participants who complete this course will acquire the knowledge to perform effective data analyses in his or her particular field of study. The student will understand the importance of effective graphical displays, the usefulness of using resistant methods in fitting models, when to use transformations to help to simplify data structures, and how to look at residuals from a model fit to look for finer structure in the data.

STUDENT OUTCOMES

Student evaluation

The grade in the class will be based on the student's performance on homework and computer labs. A typical homework assignment will consist of several data analyses following the techniques described in the lecture notes. All homeworks are to be turned in electronically or by hard-copy to the instructor using a word processor (such as Microsoft Word) using pasted output (graphs and tables) from the statistical package R. 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 scale

All homeworks will be graded on a point-system. The final grade is based on the percentage of the total points on the homework where a percentage of 90-100 is an A, a percentage of 80-89 is a B, a percentage of 70-79 is a C, a percentage of 60-69 is a D, and a percentage lower than 60 is an F. A student's participation in the course by email or postings will help determine grades in borderline cases.

COURSE OF STUDY

The calendar of events for the course, including the homework assignments, is shown below.

Week 1	Introduction to EDA, working with a single batch displays HOMEWORK: HW0, Fathom introduction
--------	---

Week 2	Working with a single batch, summaries HOMEWORK: HW 1, Fathom outliers
Week 3	Comparing batches HOMEWORK: Fathom bins
Week 4	Spread vs level plot HOMEWORK: HW 2, Fathom comparing batches
Week 5	Transformations HOMEWORK: Fathom spread vs. level
Week 6	Reexpressing for symmetry HOMEWORK: Fathom symmetry HW 3
Week 7	Plotting/Resistant lines HOMEWORK: Fathom fitting lines
Week 8	Straightening HOMEWORK: HW 4, Fathom straightening
Week 9	Smoothing sequences HOMEWORK: HW 5
Week 10	Two-way tables, median polish
Week 11	Multiplicative and extended fits HOMEWORK: HW 6
Week 12	Working with binned data HOMEWORK: HW 7
Week 13	Fraction data HOMEWORK: HW 8
Week 14	HOMEWORK: Project
Week 15	HOMEWORK: Project
Week 16	HOMEWORK: Project

Last revised August, 2006