SYBB 310: Healthcare Data Analytics in R

Course Syllabus

Fall 2014

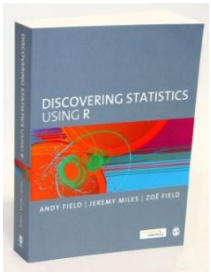
Course Description

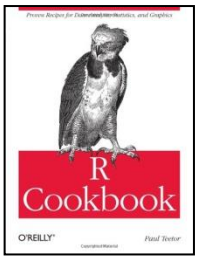
As part of the Data Science Minor, SYBB 310 is designed to introduce students to the basic tools used in data science, focusing on elementary statistics and building up to regression models. In this course, we will provide hands-on training in statistical programming through the use of the open-source statistical computing language, R. Over the semester, students will gain a practical understanding of the essential statistics needed for data science, and students will apply these principles using R to analyze a large dataset of 10,000 patients' de-identifed electronic medical records. No background in statistics or programming is expected for this course.

Undergraduate Prerequisites: EECS 131; or EECS 132; or equivalent proficiency

Undergraduate Enrollment (max): 50

Textbooks





1. Field, A., Miles, J., and Field, Z. "Discovering Statistics Using R." SAGE (2012).
2. (Optional) Teetor, P. "R Cookbook." O'Reilly (2011).

Goals

1. To provide students with a hands-on feel for practical statistics
2. To teach students how to understand and discuss inferential statistics to data
3. To teach students statistical programming and basic graphics in the R environment

Course Structure

Lectures on statistics and data science concepts will be provided in class, and attendance is expected. Apart from assigned homeworks, students are also expected to participate in an online discussion forum on data science (outlined below). *Special topics*: Time permitting, the instructor may elect to cover additional topics in data science, e.g. clustering or machine learning, in the course.

Homework

Homework will be assigned at the end of each week, and will be due within one week. Each assignment will require the student to apply the statistics & code learned that week to a real-world dataset. Each student will be expected to turn in a 3 page project report detailing their methods, results, and analysis of the homework objectives. As the methods learned each week vary significantly, the contents of the reports (e.g. types of figures to include) will differ. Thus, the instructor will provide a detailed rubric for each report upon assignment.

Online Discussion Forum

To engage students in a practical understanding of statistics and its application to data science, a portion of the class will revolve around a weekly online discussion of data science and the process of data mining. Topics can include: advanced data visualization in R; understanding user/client needs in approaching data analytics as a service; and the identification of questions where data can provide answers. At the beginning of the semester, each student will be assigned a week on which they are to turn in a 3-4 paragraph essay, in which they identify their question & its context, present the pertinent issues, and suggest solutions. Once the student has submitted his/her analysis, the discussion will be posted to the course website, where students will be expected to engage in an open-ended discussion of the aforementioned questions. Participation in this online discussion will count towards the course grade.

Grading

Grades will be assigned on the traditional scale (90-100=A, 80-89=B, etc.). However, grades may be curved at the end of the semester at the instructor's discretion.

Item	Percent
Attendance	10%
Homeworks	25%
Midterm exam	20%
Final exam	25%
Online forum: participation	10%
Online forum: analytical essay	10%
Total	**100%**

Topic Schedule

Week	Dates	Topic	Readings: from Field book	Homework using PracticeFusion data *(subject to change)*
1	Aug 25-29	Intro to Statistics • Measurement & inference • Validity & reliability • Frequencies & distributions Intro to R • commands, objects, functions • packages	Ch. 1: all Ch. 3: 3.1-3.4 Ch. 2: 2.1-2.5	• Install R • Write an R function to change the working directory and print a list of files
2	Sep 2-5 *(Sep 1: Labor Day)*	Probability distributions • Populations • Means & variance • Standard error & confidence intervals Intro to R • Manipulating data in R • Vectors, lists, arrays, data frames	Ch. 3: 3.5-3.9 Ch. 4: all	• Which distribution could be used to model blood glucose in our population? • Calculate the mean & variance for blood glucose and neutrophil count
3	Sep 8-12	Hypothesis testing • Tests of normality • Tests of variance • QQ plots Exploring dat • ggplot2 package • histograms, box plots • density plots • Data transformation & normalization	Ch. 2: 2.6	• QQ plot of blood glucose • Test the distribution for normality • Plot histogram of log-neutrophil count with overlaid normal distribution
4	Sep 15-19	Testing means • Z-tests • t-tests • Sampling distributions • Standard error	Ch. 9: all	• Test whether the average blood glucose level is significantly different between users and non-users of paroxetine
5	Sep 22-26	Covariance & Correlation • Correlation Coefficients • Tests of significance • Partial correlation	Ch. 6: all	• Plot the correlation between systolic blood pressure and serum creatinine over time & cross-sectionally • Test if this correlation is significant
6	Sep 29-Oct 3	Linear regression	Ch. 7: all	• Calculate a linear regression

		• Least squares • Goodness of fit • Diagnostics in R • Multiple regression		of the first measured serum LDL against statin prescription, adjusting for age, gender, and race
7	Oct 6-10	<u>Logistic regression</u> • Odds ratios • Prediction • Model checking • Reporting results	Ch. 8: all	• Perform a logistic regression analysis to model which patients are prescribed a statin
8	Oct 13-17	**In-class lab, review, & Midterm exam**	--	
9	Oct 20-24	<u>Categorical data analysis</u> • Pearson's chi-square & Fisher's exact tests • Likelihood ratios • Loglinear analysis	Ch. 18: all	• Use the chi-square statistic to test for significant adverse drug events
10	Oct 29-31 *(Oct 27-28: Fall Break)*	<u>Comparing means: ANOVA</u> • ANOVA as regression • Assumptions	Ch. 10: all	• Test to see if the number of statin prescriptions is significantly different among practices in Ohio
11	Nov 3-7	<u>Analysis of covariance: ANCOVA</u> • Assumptions • Implementing in R • Plots for ANCOVA	Ch. 11: all	• Are levels of LDL comparable across different types of statins? Different demographics?
12	Nov 10-14	<u>Generalized Linear Models</u> • Repeated measures • Clustered data • Calculating effect sizes	Ch. 13: all	• Calculate a linear regression of the repeatedly measured white blood cell count against antibiotic prescription, adjusting for age, gender, and race
13	Nov 17-21	<u>Multilevel Linear Models</u> • Fixed effects vs random effects • Assessing fit • Developing models in R	Ch. 19: all	• Perform a multilevel regression model to analyze how insurance status & practice type affect the brand of medication prescribed
14	Nov 24-26 *(Nov 27-28: Thanksgiving)*	Sample size & power calculations	--	• Calculate the sample size needed to detect a difference in blood glucose of 10 mg/dL at a power of 0.80
15	Dec 1-5	**In-class lab & review**	--	
16	Dec 9-17	**Final Exam**		

Dataset

Dataset for Fall 2014	**Description**
PracticeFusion Prediction Challenge	Practice Fusion is an Electronic Health Record (EHR) manufacturer, with more than 170,000 medical professional users treating 34 million patients in all 50 states. Practice Fusion's EHR-driven research dataset is used to detect disease outbreaks, identify dangerous drug interactions and compare the effectiveness of competing treatments. In partnership with Kaggle, Practice Fusion released 10,000 de-identified, HIPAA-compliant medical records to spur innovation into new uses of clinical data to improve public health and patient care. This dataset is one of the largest and richest sources of medical record data ever released and includes information on diagnoses, lab results, medications, allergies, immunizations, vital signs, and health behavior.