Data Analysis and Visualization

CS555

Kia Teymourian
Email: kiat@bu.edu
Office hours: Thursday afternoon 2-5pm or by appointment
Office Location: 808 Commonwealth Avenue, Room 257

Course Description

This course provides an overview of the statistical tools most commonly used to process, analyze, and visualize data. Topics include describing data, statistical inference, 1 and 2 sample tests of means and proportions, simple linear regression, multiple regression, logistic regression, analysis of variance, and regression diagnostics. These topics are explored using the statistical package R, with a focus on understanding how to use and interpret output from this software as well as how to visualize results. In each topic area, the methodology, including underlying assumptions and the mechanics of how it all works along with appropriate interpretation of the results, are discussed. Concepts are presented in context of real world examples.

Learning Objectives

By successfully completing this course you will be able to:

- Appreciate the science of statistics and the scope of its potential applications
- Summarize and present data in meaningful ways
- Select the appropriate statistical analysis depending on the research question at hand
- Form testable hypotheses that can be evaluated using common statistical analyses
- Understand and verify the underlying assumptions of a particular analysis
- Effectively and clearly communicate results from analyses performed to others
- Conduct, present, and interpret common statistical analyses using R

Books

The following two books are required for the course. These should be used as reference material to help support you in your assignments and supplementing the course's Live Classroom sessions on R. The modules themselves will provide you with the necessary information for the theory, concepts, and examples that you will need to complete your quizzes and understand the methodologies that you will apply to the problems presented in the homework assignments.

There will be no reading assignments from these books. These are excellent supplemental texts that you may want to review as we go through the course and also keep as reference text as you continue to use R in the future.

- Chang, W. (2013). R graphics cookbook. Sebastopol, CA: O'Reilly. ISBN 9781449316952.
- Teetor, P. (2011). R cookbook. Sebastopol, CA: O'Reilly. ISBN 9780596809157.

Additional Text Books for further reading:

- https://www.openintro.org/stat/ Free PDF for download & R tutorials and codes.
- Andy Field, Jeremy Miles and Zoe Field. (2012) Discovering Statistics Using R. Publisher: SAGE Publications Ltd. ISBN-13: 978-1446200469
- Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. (2013) An Introduction to Statistical Learning with Applications in R. Springer.

Class Policies

1) Attendance & Absences – Full attendance and participation is expected. If there is a reason to miss a session, advanced notice through email should be sent to the lecturer.

2) Assignment Completion & Late Work – All assignments should be submitted on time. If there is a delay, the student must be in touch with the instructor. Late submissions without reasons will result in grade deduction.

3) Academic Conduct Code –Cheating and plagiarism will not be tolerated in any Metropolitan College course. They will result in no credit for the assignment or examination and may lead to disciplinary actions.

Please take the time to review the Student Academic Conduct Code:
http://www.bu.edu/met/metropolitan_college_people/student/resources/conduct/code.html

NOTE: [This should not be understood as a discouragement for discussing the material or your particular approach to a problem with other students in the class. On the contrary – you should share your thoughts, questions and solutions. Naturally, if you choose to work in a group, you will be expected to come up with more than one and highly original solutions rather than the same mistakes.]

Grading Criteria

Homework Assignments

The six homework assignments are focused on applying theory learned in the week's module to a set of data and analyzing that data in R. Assignment submissions should be a single Microsoft Word or PDF file. The R code used to generate your results should be appended to the end of your assignment.

Lectures relating to R will be held and recordings will be posted after each session. Slides from the session will also be made available to students.

Quizzes

The six quizzes will evaluate students understanding of concepts presented in the corresponding week's module. Students should ensure adequate preparation before starting the quiz. It will not be possible to do well on the quiz without first reviewing the course material in depth and attempting to understand all examples and test yourself questions. It is recommended that you complete the quiz after you feel comfortable with the material and asked any questions that you may have had.

Final Examination

The final exam will be comprehensive and will cover material from the entire course. It will be an open-book proctored exam consisting of questions similar to the ones in the quizzes but longer in length.

The final grade for this course will be based on the following:

Deliverable	**Weight**
Homework Assignments	40%
Quizzes	30%
Final Exam	30%

Study Guide

Lecture 1 Introduction to the science of statistics - part 1

- Fundamental Elements of Statistics
- Qualitative and Quantitative Data Summaries

Lecture 2 Introduction to the science of statistics - part 2

- Normal distribution
- Sampling
- The Central Limit Theorem

Lecture 3 Confidence intervals and hypothesis tests - part 1

- Statistical Inference
- Stating Hypotheses
- Test Statistics and p-Values
- Evaluating Hypotheses

Lecture 4 Confidence intervals and hypothesis tests - part 2

- Significance Test “Recipe”
- Significance Tests and Confidence Intervals
- Inference about a Population Mean
- Two-Sample Problems

Lecture 5 Understanding the association between two continuous or quantitative factors - part 1

- Scatterplots
- Correlation

Lecture 6 Understanding the association between two continuous or quantitative factors - part 2

- Simple Linear Regression
- F-test for Simple Linear Regression
- t-test for Simple Linear Regression

Lecture 7 Regression diagnostics

- Residual Plots
- Outliers and Influence Points
- Assumptions of least-square regression

Lecture 8 Multiple linear regression

- Equation of multiple linear regression
- Interpretation of multiple linear regression
- F-test for Multiple Linear Regression
- t-tests in Multiple Linear Regression
- Cautions about Regression

Lecture 9 Analysis of Variance (ANOVA) - part 1

- One-Way Analysis of Variance
- F-test for ANOVA
- Evaluating Group Differences
- Type I and Type II Errors

Lecture 10 Analysis of Variance (ANOVA) - part 2

- Issues with Multiple Comparisons
- Assumptions of Analysis of Variance
- Relationship between One-Way Analysis of Variance and Regression
- One-Way Analysis of Covariance
- Two-Way Analysis of Variance
- Two-Way Analysis of Covariance

Lecture 11 Analysis for proportions - part 1

- One-Sample Tests for Proportions
- Significance Tests for a Proportion

- Confidence Intervals for a Proportion

Lecture 12 Analysis for proportions - part 2

- Two-Sample Tests for Proportions
- Confidence Intervals for Differences in Proportions
- Significance Tests for Differences in Proportions
- Effect Measures
- Logistic Regression
- Multiple Logistic Regression
- Area under the ROC Curve

Lecture 13 Review session