Mathematics 156/E-156, Spring 2016
Mathematical Foundations of Statistical Software

Final Project Guidelines

Last Modified: April 20, 2016

Project is due May 4th, 2016 at Noon
1 Point per hour will be deducted for each hour (rounded up) that the project is late

Required technical elements – the dataset – 4 points

1. A dataframe
2. At least two categorical or logical columns (ie factors)
3. At least two numeric columns
4. At least 20 rows, preferably more, but real-world data may be limited

Required technical elements – analysis – 4 points
Any two of the following (2 points each)

1. Linear regression
2. Student t confidence interval
3. Bayesian prior updated by data

Required technical elements – graphical display – 2 points
Any two of the following

1. A scatter plot with regression line
2. A plot showing Bayesian prior and posterior distributions
3. A display illustrating confidence intervals

Required technical elements – presentation – 6 points

1. A .csv file with the dataset, uploaded to the course website
    a. Any data scraped from the internet should be saved and then accessed locally, ie, rerunning the script should not re-scrape the data
2. A long, well-commented script that loads, explores, and analyzes the data
    a. Comments should not exceed 80 characters, ie, span multiple short lines
3. A short script that presents interesting highlights in ten minutes
4. A one-page handout (bring 22 copies) that explains the dataset and summarizes the analysis
5. A one-paragraph abstract
6. Deadlines met

Points for creativity or complexity – maximum of 11

1. Use all three required analysis technical elements (2 points)
2. Comparison of analysis by classical methods and simulation methods
3. Comparison of analysis by Bayesian and frequentist approaches
4. Use of a Bayesian conjugate family beyond the two studied in class
5. Calculation and display of a logistic regression curve
6. A dataset with many (10+) columns, allowing comparison of many variables

7. A graphical display unlike one presented in the textbook or course scripts
8. Appropriate use of R functions for a distribution and its conjugate prior
9. Appropriate use of bootstrap techniques (2 points)
10. A convincing demonstration of an unexpected statistically significant relationship
11. A convincing demonstration that a relationship expected to be significant is not
12. Professional-looking software engineering (functions instead of copy-paste!)
13. Nicely labeled and formatted graphics (feel free to reach out to Stu for ggplot pointers)
14. Appropriate use of novel statistics (eg, trimmed mean, skewness, median absolute deviation, least-absolute-error regression, ratios, order statistics, R squared)
15. Use of theoretical knowledge of chi-square, gamma, or beta distributions
16. Maximum-likelihood estimation of parameters (2 points)
17. Appropriate use of covariance or correlation
18. Team consists of exactly two members (1 to 3 allowed)
19. Team includes a Harvard College student and an Extension student
20. Team includes a distance student and an on-location student
21. A video of the short script is posted to YouTube and a link to it is on the course site
22. A document of about 5 pages about the project is created within R (using Markdown, knitr, or something similar)

Subject Impression – if these folks were applying for a job that requires computerized statistical analysis, I would…

1. Immediately disband the search committee and hire them. (3 points)
2. Add them to a shortlist of leading candidates. (2 points)
3. View them as acceptable if no one better turns up. (1 point)