Group Project

Statistics 506, Fall 2018 (./index.html)

Group Project

The group project will be completed in groups of three. Groups have been assigned randomly and will be posted to Canvas. There may be 1-2 groups of four.

Each group will choose a data management, analysis, or visualization technique and produce a tutorial on the selected topic. The tutorial should include:

- A short description or explanation of the topic;
- A real data set on which to demonstrate the technique;
- Three approximate translations of the selected example (one per group member). These translation must use at least two software languages (Stata, R, SAS, etc) and at least one version must be in R. Up to two versions may be in R, provided they make use of different packages for the core technique (i.e. one could use base R and another data.tables or dplyr).

Topics should be of no greater scope than those at the UCLA page here (https://stats.idre.ucla.edu/other/dae/). Please keep other time demands in mind when deciding the scope of your project. It is preferable to have a well put together project of limited scope than a less-polished project on a broader topic.

You can find completed projects from last year's course here (https://jbhender.github.io/Stats506/F17/CourseProject.html).

Each group should write a short proposal containing:

- the topic for the tutorial
- the data to be used as an example
- the languages/packages to be used for the examples.

Each member of the group should assume primary responsibility for an example script and these responsibilities should be included in your proposal.

A group liaison should submit the group's proposal to me via email with the subject header "Stats 506 Group Project Proposal".

I plan to post your tutorials to this page – you may include or omit your name from the author information at your own discretion.

Guidelines

Introduction and overview

Your introduction and overview should be approximately 3-5 paragraphs explaining:

- what your topic is
- when it is useful and/or why it is important
- important information about the topic
- sources for the information you provide and additional resources for learning more about the topic
- a link to and brief description of the data used in the tutorial
- the scope of your tutorial
- the languages used in your tutorial
- reasons, if any, you could not obtain the same results from all three language examples.

Examples

All three examples should follow a common outline to the extent possible. Deviations should be due to limitations or stylistic differeneces in the languages chosen rather than lack of coordiation among group members. Where deviations do occur, please explain and justify them in the text surrounding the example.

It is permissible for one or more examples to extend beyond the scope of others provided that all three examples share a common core set of tasks.

You may use a language (i.e. Python, Matlab) not taught in this course for one or more examples as long as I approve it in your proposal.

Git

Groups should use git to coordinate their work. Each member of the group should create an account at github.com. One group member should create a public repository for the project with others submitting pull requests (https://help.github.com/articles/about-pull-requests/) to them. Your git repo is considered part of the final submission and should include at minimum:

- sources files for your tutorial including all examples

- Rmd files should be included
- scripts should be included
- an html page should be included
- a readme
- a minimum of two commits per team member.

Excluding extraordinary circumstances, all group members will receive the same grade. However, I reserve the right to modify this policy in cases where one or more group members clearly put in less effort than the others.

Timeline

1. **Topic Proposals due Friday November 16 at 4pm**: Your group must have your proposal approved by me prior to this time. Groups are required to select unique topics so it is to your benefit to submit early.
2. **Draft Due Date: Tuesday November 27 at 4pm**. Drafts should be mostly complete and contain a concise to-do list of outstanding items. Submit drafts to Canvas as a zip archive and link to the official git repo.
3. **Peer Review: Due Monday December 3 at 5pm**. You will be asked to provide constructive feedback to another group. Guidelines for how to structure this feedback will be provided.
4. **Final Due Date: Friday December 7 at 5pm**. This is the deadline to submit the final version of your tutorial. Please submit as a zip archive to canvas with html, Rmd, and (possibly) scripts or data files included.
   Please make edits in response to peer feedback.

Approved Group Proposals

I will post group proposals here as they are approved. Two groups may not choose the same or closely related topics. The two topics below are also reserved.

If you are repeating a topic from last year, please be clear how your tutorial will add value to what was done previously.

Reserved

1. Reshaping data between long and wide formats
2. The “split-apply-combine” pattern

Approved

1. **Group 1** -
   - Topic: Ridge Regression;
   - Data: meatspec (https://cran.r-project.org/web/packages/faraway/faraway.pdf);
   - Languages: Matlab, R, Stata
   - Tutorial (./GP/Group1.html)
2. **Group 2** -
   - Topic: Factor Analysis;
   - Data: bfi (https://www.personality-project.org/r/html/bfi.html);
   - Languages: Python (sklearn), R (psych), Stata
   - Tutorial (./GP/Group2.html)
3. **Group 3** -
   - Topic: Parametric and Non-Parametric ANOVA;
   - Data: Rat Survival (http://sia.webpopix.org/statisticalTests2.html);
   - Languages: Matlab, R, Stata
   - Tutorial (./GP/Group3.html)
4. **Group 4** -
   - Topic: Ordinal Logistic Regression;
   - Data: Soup (https://rdrr.io/rforge/ordinal/man/soup.html);
   - Languages: R (ordinal), SAS, Stata
   - Tutorial (./GP/Group4.html)
5. **Group 5** -
   - Topic: Logistic Regression with Model Diagnostics;
   - Data: Pima (https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/Pima.tr.html);
   - Languages: Python (sklearn, pysal), R, Stata
   - Tutorial (./GP/Group5.html)
6. **Group 6** -
   - Topic: Truncated negative binomial regression;
   - Data: Abalone (https://archive.ics.uci.edu/ml/datasets/abalone);
   - Languages: R, SAS, Stata
   - Tutorial (./GP/Group6.html)
7. **Group 7** -
   - Topic: Clustering using Finite Mixture Models;
   - Data: Wine (https://archive.ics.uci.edu/ml/datasets/wine);
   - Languages: Python, R, Stata
   - Tutorial (./GP/Group7.html)
8. **Group 8** -
   - Topic: Fixed effects models;
   - Data: Cigar (https://rdrr.io/rforge/plm/man/Cigar.html);
   - Languages: R (plm), SAS (proc reg), Stata (xtreg, areg, reghdfe)

- Tutorial (./GP/Group8.html)

1. **Group 9** -
   - Topic: Cubic splines regression;
   - Data: uswages (https://rdrr.io/cran/faraway/man/uswages.html);
   - Languages: Python (statsmodels), R (splines), Stata (mkspline)
   - Tutorial (./GP/Group9.html)
2. **Group 10** -
   - Topic: (Divisive) Hierachical Clustering;
   - Data: black-friday (https://www.kaggle.com/mehdidag/black-friday);
   - Languages: Python (sklearn.cluster), R (hclust), Stata (cluster)
   - Tutorial (./GP/Group10.html)
3. **Group 11** -
   - Topic: Linear Discriminant Analysis (LDA);
   - Data: Seeds (https://archive.ics.uci.edu/ml/datasets/seeds);
   - Languages: R, SAS, Stata
   - Tutorial (./GP/Group11.html)
4. **Group 12** -
   - Topic: Multinomial Logistic Regression (all vs reference);
   - Data: iris;
   - Languages: R (multinom), SAS (proc logistic), Stata (mlogit)
   - Tutorial (./GP/Group12.html)
5. **Group 13** -
   - Topic: Comparing logit and probit regression;
   - Data: Communities and Crime (https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime);
   - Languages: Python (sklearn, pysal), R, Stata
   - Tutorial (./GP/Group13.html)
6. **Group 14** -
   - Topic: Probit Regression;
   - Data: Mroz (http://www.econometrics.com/comdata/wooldridge/data.html);
   - Languages: Python (StatsModels), R (glm), SAS, Stata
   - Tutorial (./GP/Group14.html)
7. **Group 15** -
   - Topic: (Agglomertive) Hierarchical Clustering using Average Linkage;
   - Data: USArrests (https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/USArrests.html);
   - Languages: Python (sklearn, scipy), R (cluster), Stata (cluster)
   - Tutorial (./GP/Group15.html)
8. **Group 16** -
   - Topic: Linear Mixed Effects Models;
   - Data: PM2.5 (https://archive.ics.uci.edu/ml/datasets/PM2.5+Data+of+Five+Chinese+Cities);
   - Languages: Python (Statsmodels), R (lme4 and glmmADMB)
   - Tutorial (./GP/Group16.html)
9. **Group 17** -
   - Topic: Multidimesional Scaling;
   - Data: 93cars (http://jse.amstat.org/datasets/93cars.dat.txt?fbclid=IwAR19EJDNQzMMQOBe6ZC4t4li0zx0UpkDBA0SWDNWa3T54JXc-yfpPiPIqsk)
   - Languages: Stata, R (dplyr + cmdscale, data.table + smacof)
   - Tutorial (./GP/Group17.html)
10. **Group 18** -
    - Topic: Multilayer Perceptron Models;
    - Data: Telco Customer Churn (https://www.kaggle.com/blastchar/telco-customer-churn);
    - Languages: Python (tensorflow and base), R (MXNet)
    - Tutorial (./GP/Group18/introduction.html)
11. **Group 19** -
    - Topic: Model Selection and Diagnostics in Linear Regression;
    - Data: Insurance (https://www.kaggle.com/mirichoi0218/insurance);
    - Languages: Matlab, Python, R
    - Tutorial (./GP/Group19.html)
12. **Group 20** -
    - Topic: Canonical Correlation Analysis
    - Data: NHANES 2003-2004 (https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2003)
    - Languages: R, SAS, Stata; Reference: Liu et al 2009 (https://www.ncbi.nlm.nih.gov/pubmed/?term=Examination+of+the+relationships+between+environmental+exposures+to+volatile+organic+compounds+and+biochemical+live
    - Tutorial (./GP/Group20.html)
13. **Group 21** -
    - Topic: Monte-Carlo simulation of Portfolio Stock Returns;
    - Data: Stock prices for Apple, Facebook, and Google 11/14/2017 - 11/14/2018;
    - Languages: Matlab, Python, R
    - Tutorial (./GP/Group21.html)