Stat 343 - Project Description and Outline

Overview

In this project, you will work in a group learn about and implement a statistical model or estimator that we have not discussed in class. You will conduct a simulation study and apply the method or model to a real data set. You will also develop a short written project report and give a brief oral presentation describing your work.

Group Formation

You will work in a group of two or three, of your choosing (I may consider groups of size two or four in some cases). Each group will be assigned a letter, and you will use "GroupX" (note capitalization and no spaces) with your particular letter in the subject line of all communications with me from that point forward.

Example Topics

Although we have explored a variety of new statistical models in this class and have developed maximum likelihood and Bayesian inference procedures in a general sense, the world of statistics is very large. In this project, you will have a chance to explore a new topic in more depth. Below are a few suggestions for possible project topics; you can select from among these or propose another topic. I am always happy to help with pointers towards appropriate sources and technical details.

- **Inference from Complex Survey Data**. Many surveys and opinion polls sample individuals from the population with different weights (for instance, political polls often stratify according to variables such as race, and sample from the different strata with different probabilities). In this project, you would implement methods for estimating a population parameter (such as a mean or proportion) using survey data with unequal weights. Chapter 7 (especially Section 7.6) of Rice discusses these methods in some depth. Another more general approach is the Horvitz-Thompson estimator, which can be used to analyze more general survey data where different people in the population had different probabilities of being included in the sample. Many government data sets, and polling data sets from the Pew Research Center, include sampling weights for the subjects.
- **Gibbs Sampler or Metropolis-Hastings Sampler**. We briefly described the Metropolis-Hastings sampler, which can be used to perform Bayesian inference via MCMC. In this project, you would pick a model with at least two parameters and hand-code either a Gibbs sampler or a Metropolis-Hastings sampler to perform inference for that model. "Introduction to Statistical Thought," by Michael Lavine, is a freely available textbook with a readable introduction to Metropolis-Hastings including example code in R. Gibbs is briefly discussed there, as well as in Rice and DeGroot and Schervish.
- **Kernel Density Estimation**. Kernel Density Estimation is a non-parametric approach to density estimation. This is the method for producing smoothed histograms that is behind the `density` function in base R and the `geom_density` geometry type in ggplot2. In this project you would learn about Kernel Density Estimation for estimating the distribution of a univariate random variable. A Kernel Density Estimate depends on a parameter called the *bandwidth*, which is similar to the bin width of a histogram and controls the smoothness of the density estimate. In this project, you would implement estimation of this bandwidth parameter via cross-validation. Good informal places to start reading about the method are at http://www.mvstat.net/tduong/research/seminars/seminar-2001-05/ and https://en.wikipedia.org/wiki/Kernel_density_estimation. I would guide you through performing cross-validation for estimation.
- **Mixture of Normals; E-M Algorithm**. Another flexible model for the distribution of a random variable is a mixture of normals. In this model, the distribution is represented as a weighted combination of normal distributions with different means and variances. For example, this model can be helpful if the distribution of a variable is multimodal. In this project, you would learn about and implement the Expectation-Maximization (E-M) algorithm for estimating the means and variances of the normal mixture components, as well as the weight assigned to each mixture component. A detailed discussion of this is in DeGroot and Schervish.
- **Generalized Linear Models**. A generalized linear model is a common approach to modeling the relationships among multiple variables when a linear regression model with normally-distributed residuals is not reasonable. You could implement estimation for a Generalized Linear Model in a Bayesian or frequentist paradigm. Some example implementations in a Bayesian framework can be found at https://github.com/stan-dev/example-models/wiki/

ARM-Models-Sorted-by-Type. Note that it would not be valid to simply re-use one of those analyses; however, it would be valid to do a detailed write-up of the model and the code, apply that model to a new data set and evaluate estimation performance with a simulation study.

This is not intended to be an original research project. All of the ideas outlined above are well-established methods that you have the necessary background to learn about and implement within a few weeks (at least, a basic version of the model).

There are many, many other possibilities out there for good projects. If there's a type of model you're interested in, or even if there is a general data analysis problem you're interested in but you don't know what methods are used in that context, let me know! I will try to help identify the relevant statistical methods and a feasible project in that area. The only restriction is that the project topic should be beyond the scope of standard regression techniques that are covered in more depth in other courses in the statistics curriculum.

General Rules

The project deliverables will include:

- an oral presentation approximately 15 minutes in length (everyone in your group must speak)
- an accompanying handout (front and back of a single piece of paper) describing your project at a high level
- a report written as a Jupyter notebook or R Markdown document

You will need to consult other sources for information about the topic of your project, and you should credit these sources in your report.

Feel free to consult me about any statistical or programming questions that arise.

Timeline

Checkpoint	Due Date	Credit	Submission Method
Group roster	3/22	2 pts	Email with subject line "Stat 343 Roster"
Project proposals	3/27	5 pts	Push to GitHub
Revised project proposal	4/3	5 pts	Push to GitHub
Detailed method and data set description	4/10	5 pts	Push to GitHub
Method implemented in R; functioning proof of concept with at least one of data application and simulation study complete	4/17	5 pts	Push to GitHub
Oral presentation	4/24, 4/26,& 4/27	50 pts	In class
Project report (technical appendix)	5/7	50 pts	Push to GitHub
Group Dynamic	5/7	5 pts	Email with subject line "Stat 343 GroupX Dynamic"

Roster

Form a group of 3 students. Have one person send the group roster electronically by the date listed above, with appropriate cc's, using the message subject header "Stat 343 Roster".

Project Proposals

You must submit proposals for at least two projects your group is interested in to me, along with a ranking of which of these projects you are most excited about. Each project proposal should include the following:

- A fairly detailed description of the statistical problem that you will address and a statement of the methods you will explore for solving that problem. This will look similar to the brief descriptions I gave of sample topics in the "Assignment Topic" section above, but will include slightly more detail. For example, in the "Gibbs Sampler or Metropolis-Hastings Sampler" topic, I would want you to have at least some general idea of what model you will be fitting. If you need help figuring out the details, be in touch!
- One or two concrete ideas of potential example applications to real data. You don't need to have downloaded the data, but you should include a couple of links to places where you are confident that you can find relevant data sets. Again, if you need help finding appropriate data, let me know.

Revised Project Proposal

I will give you feedback and comments on your initial project proposals by the end of the day on 3/29. These will include suggestions to either scale back the scope of the project or add a new piece to ensure that the project is at an achievable but still interesting level; suggestions for further reading; and any suggestions that I have for what your model, estimation strategy, simulation study, or application to data should look like.

In your revised project proposal, you will incorporate these suggestions into your project proposal, create an outline of your final report, and start filling in some of the details. This should include a first attempt at writing down the following:

- model structure including distributions for the data and any prior distributions
- a rough description of how estimation is to be performed
- a rough description of your simulation study, including:
  - How many simulations you plan to conduct
  - How data will be simulated
  - How you will measure how well your method performs
- a first look at a data set

It's ok if there are errors, mistakes, or points of confusion in any of the above at this point. Also, as long as you have picked out a data set and at least made an attempt to read it in, it's ok if you haven't done much with the data yet.

Detailed method and data set description

Your goal at this point is to be confident in your understanding of the statistical method you are using, so that you can implement the method. You should have written up a detailed description of the method and estimation procedures, possibly with a graphic or two to help illustrate it.

Additionally, I would like to see a final description of the data set you are working with, including the data source, the variables that will be used in your example, and plots.

Method implemented in R

You should have working code to perform estimation, and a demonstration that this code works on at least one of the application to data and the simulation study.

Oral presentation and project report

In your oral presentation and project report, you will present an overview of the statistical problem you are solving, the method or model you are using and how it works, how estimation was performed, and the results of the simulation study and application to data. Your project report will be mostly written by the previous check-in.

Group Dynamic Report

Ideally, all group members would be equally involved and able and committed to the project. In reality, it doesn't always work that way. I'd like to reward people fairly for their efforts in this group endeavor, because it's inevitable that there will be variation in how high a priority people put on this class and how much effort they put into this project.

To this end I will ask each of you (individually) to describe how well (or how poorly!) your project group worked together and shared the load. Also give some specific comments describing each member's overall effort. Were there certain group members who really put out exceptional effort and deserve special recognition? Conversely, were there group members who really weren't carrying their own weight? And then, at the end of your assessment, estimate the percentage of the total amount of work/effort done by each member. (Be sure your percentages sum to 100%!)

For example, suppose you have 3 group members: X, Y and Z. In the (unlikely) event that each member contributed equally, you could assign:

- 33.3% for member X, 33.3% for member Y, and 33.3% for member Z

Or in case person Z did twice as much work as each other member, you could assign:

- 25% for member X, 25% for member Y, and 50% for member Z

Or if member Y didn't really do squat, you could assign:

- 45% for member X, 10% for member Y, and 45% for member Z

I'll find a fair way to synthesize the (possibly conflicting) assessments within each group. And eventually I'll find a way to fairly incorporate this assessment of effort and cooperation in each individual's overall grade. Don't pressure one another to give everyone glowing reports unless it's warranted, and don't feel pressured to share your reports with one another. Just be fair to yourselves and to one another. Let me know if you have any questions or if you run into any problems.

Assessment Criteria

- General: Is the topic interesting and substantial? This project is not intended to be an original research project; at the same time, I don't want you to simply reproduce someone else's analysis. During the project proposal phase, I will endeavor to make sure you are all set up with a project that it is at the appropriate level.
- Technical Mastery: Do you demonstrate that you understand the methods you are using? Does the submitted R code work correctly?
- Written Report: How effectively does the written report communicate the goals, procedures, and results of the study? Are the claims adequately supported? How well is the report structured and organized? Are all of the figures and tables numbered, captioned and appropriately referenced? Does the writing style enhance what the group is trying to communicate? How well is the report edited? Are the statistical claims justified?
- Oral Presentation: How effectively does the oral presentation communicate the goals, procedures, and results of the study? Do the slides help to illustrate the points being made by the speaker without distracting the audience?