Final Project

Loy, Math 315

Due by 3:00 p.m., Monday, November 25

The final project is an opportunity to use the methods learned in this course to address a question of your choice through data analysis, or to explore a new method in Bayesian statistics. While we will not cover all of the model types in the textbook, you could certainly learn about one of the models we don't discuss and use that in your project. For example, if we don't get to Poisson regression, you can still think about how to set up the model in JAGS and interpret the results; or, if you have multilevel data (such as students within schools) you could read enough about hierarchical models to implement one.

Working in groups: You should work in groups of up to three people. You will need to fill out the Google form indicating who is in your group. Only one person from each group needs to so this. If you need help finding a partner, you still should fill out the Google form so that I can match you with a group.

The expectation is that each student shares equal responsibility for the report. In a group project, you will turn in one report but you will also each separately assess the contribution made by each group member. Usually, students working in groups will receive the same grade on the report, but on the basis of these assessments I have the discretion to award different grades to each student.

Working in individually. If you would like to work individually, then you will need to email a petition explaining why you need to work individually.

Deadlines:

- Monday, 4 Nov. Confirm groups. Fill out this Google form (only one group member needs to do this): https://forms.gle/w9AGeQsZ5ARGWAR29.
- Monday, 11 Nov. Submit your proposed data set and a brief proposal by 4 p.m. Your proposal should describe the general topic/phenomenon you want to study, as well some focused questions that you hope to answer and specific hypotheses that you intend to assess. You should also provide an overview of your data set and how it links to your goals. Be sure to clearly communicate the source of your data set.
- Monday, 25 Nov. Final project due by 3:00 p.m. Email a .pdf version of your paper to me.
- Monday, 25 Nov. Email the supplemental materials to me including: (1) your data (in a .csv file), (2) your code supplement (in a .Rmd or .R file). Your code supplement should be organized and clearly commented. If for some reason your data file is too large, feel free to share it via Dropbox or Google Drive.
- Monday, 25 Nov. Complete the assessment of your contribution to the project and that of your group by midnight. The link to this Google form will be posted on GitHub. Failure to complete this assessment will result in a 5% reduction of your project grade. Further, I will use this feedback and adjust individual grades if there appeared to be a problem sharing the workload.

Data

Start by considering what your cases of interest are (for example: people, cities, plots of land, restaurants, a particular plant, etc) and how you will collect data that satisfies the inference conditions discussed in class. Your data set must be appropriate for the proposed model, and should contain at least two predictor variables. You may not use a data set that from textbooks.

Report structure

The report must be organized into the following clearly labeled sections. "Clearly labeled" means that the report has bold section titles with the names listed below.

1. Title and authors. What is your report about and who is it by?
2. Abstract. A brief, one-paragraph executive summary of the problem and your findings (200 words or fewer).
3. Introduction. Describe the problem being studied / what question is being addressed. Give relevant background information as appropriate (such as information from previous studies).
4. Data. Describe how your data were collected (list your sources and describe methods). What are the cases? What are the variables? Include any information or definitions that help establish the context of your data. You should include information like the range of values for each variable (including units), the relationship between the variables, the number of observations, etc. This section can include tables or figures to help the reader understand the data.
5. Model. Introduce the model that you are fitting to the data using statistical notation. Your discussion should include how the parameters in this model will be estimated and how this model will address the scientific question of interests. Any notation you introduce, e.g. Y_i, X_i, etc., should be defined. Statistical notation does not need to be defined—for example, i.i.d. and $\mathcal{N}(\mu, \sigma^2)$—but conventions used in this class should be followed. If there is any doubt about the convention, then it is best to be explicit—for example, you can specify whether the second parameter of the normal distribution is a standard deviation, a variance, or a precision.
6. Computation. Provide the details of how you fit your model. I expect all of the projects will depend on MCMC. In this case, you should provide details of the MCMC, e.g. length of burn-in, the number of posterior draws, the number of chains, as well as an thorough assessment of convergence.
7. Results. Report on your statistical analysis in writing. Summarize the posterior distribution of the model parameters. This is typically best done by plots of posterior distributions or tables of means, standard deviations, and intervals. The answer to the scientific question should be clearly presented.
8. Discussion. Interpret the results you presented in the previous section. It also provides an opportunity to discuss any shortcomings in the model or the data, or ideas for future exploration.
9. References. Cite sources (including data sources), if applicable. In statistics, we prefer the APA citation style.

10. Appendix (optional). Any supplemental tables, graphs, or analyses that are not central to your main narrative but that you found interesting.

Report write-up guidelines

- Report must be typed, double-spaced, with pages numbered and figures inserted neatly into the narrative.
- Graphs must be report quality: properly captioned, axes labeled, etc.
- Please resize graphs so they do not take up an entire page (i.e. I do not want to see a large graph with a lot of white space below).
- Do not use variable names in the report. (e.g. "The mean of the weight adult cats is...," **not** "The mean of CatWt is...").
- Be sure to include your model equations (or clear statements of your model) and posterior summaries.
- Your analyses should be stated in context. For instance, "We are 95% certain that the odds of a male dying from a heart attack is from 1.5 to 2.3 times higher than a female dying, controlling for age, use of alcohol and smoking status.
- No R output should be included (except graphs). Output from R commands should be put neatly in properly formatted tables.
- The paper should be approximately 8-12 pages.
- You will be graded on the depth, quality and clarity of your work as well as on the correctness of the statistical analyses. Clarity includes writing style, grammar, and mechanics.