Final project

Report due Wednesday, May 1st, 2024
Presentations on April 24th and 29th
PUBH 8442: Bayes Decision Theory and Data Analysis

For this project you are to conceive of, estimate, and interpret the results of a Bayesian model using real data from a chosen application area. You will summarize your analysis in a brief presentation of approximately 8-10 minutes in class on April 24th or April 29th (*5 pts*). You will also describe your analysis in a brief report, to be turned in by May 3rd (*15 pts*) by 5 pm on Canvas. The text of the report should be 2-4 pages, in addition to any relevant figures and code (*R* or *WinBUGS*) used for estimation. In your report you may incorporate the following topics:

- Description of data and analysis goal(s).
- Description of your prior and sampling model.
- Description of your estimation procedure (for example, overview of MCMC sampling algorithm).
- Assessment of your estimation procedure – does the estimation approach seem to provide a good approximation to the true posterior?
- Assessment of your model – how well does the model fit your data? Is it superior to other possible models?
- Results and conclusions from your analysis (this need not be too scientific, simply interpreting the statistical results in plain English will suffice).

You are free to find and analyze your own data (or any other dataset you like) for this project. Alternatively, you may use one of the datasets below. If you do not use one of the datasets below, please clearly explain how you obtained your data in your report.

1. *Beta-Blocker Clinical Trials*
   These are data for multiple clinical trials on the effectiveness of beta-blockers after a patient has a heart attack (myocardial infarction). Data are given for 22 independent clinical studies, each involving a different set of participants. For each study the number of deaths among participants not given beta blockers (control) and those given beta blockers (treated) is provided, as well as the total number of participants within each group.
   Data: http://www.stat.columbia.edu/~gelman/book/data/meta.asc
   Original source: http://www.ncbi.nlm.nih.gov/pubmed/2858114

2. *Cancer Gene Screening*
   These are gene expression microarray data for 645 genes, from 348 breast cancer tumors from different individuals. For each gene (row) and sample (column) the given value is a measure of how active that gene is ("expression") for that sample. The clinical subtype for each sample is given as "Basal" or "non-Basal". Basal tumors have a poorer prognosis and respond to certain therapies differently; we are interested in which genes show differential expression in Basal tumors.
   Data: An R file with the matrix of expression values, subtype label for each column, and

gene name for each row is available at http://www.ericfrazerlock.com/TCGA_Breast_Data.Rdata
Original source: http://www.nature.com/nature/journal/v490/n7418/full/nature11412.html

3. *Air pollution*
Air population data for 41 US cities, with the following variables:
City: City
SO2: Sulfur dioxide content of air in micrograms per cubic meter
Temp: Average annual temperature in degrees Fahrenheit
Man: Number of manufacturing enterprises employing 20 or more workers
Pop: Population size in thousands from the 1970 census
Wind: Average annual wind speed in miles per hour
Rain: Average annual precipitation in inches
RainDays: Average number of days with precipitation per year.

Consider, for example, a predictive model for SO2 content based on other variables.
Data: http://www.ericfrazerlock.com/AirData.csv
Original source: http://math.fau.edu/Qian/course/sta4234/airpolut.htm

4. To easily collect and obtain other datasets on TV series, used cars, or property listings, see http://myslu.stlawu.edu/~clee/dataset/ .

5. For data on COVID-19 cases, hospitalizations, and deaths (with prognostic modeling estimates) over time for several countries see https://www.healthdata.org/covid/data-downloads.