Applied Statistical Modeling and Inference

Final Project, due Sunday, May 15th, 9:00 PM
Note: this due date is not flexible. Late work will not be accepted.

The goal of this project is to explore various modeling options surrounding meta-analysis, and to communicate the implications of these choices to a general scientific audience. Often, the role of a statistician/data analyst/data scientist on any research or work team is to explore various statistical options, and also to communicate the implications of these modeling choices to others involved. Your task is to write a report that goes through four modeling options: frequentist fixed-effects meta analysis, frequentist random-effects meta analysis, Bayesian reference prior, and Bayesian skeptical prior, and explains the relative benefits of each modeling choice.

- First, read the paper entitled "Being skeptical about meta-analyses: a Bayesian perspective on magnesium trials in myocardial infarction" by Julian Higgins and David Spiegelhalter.
- You should feel free to consult (but not directly copy!) the multitude of resources on the internet about meta-analysis, including those on this well-known dataset. However, you must cite any sources that you use in your project.
- Submit a written report as outlined below. Your report should be written to a general statistical audience.
- Note that you will be graded both on the content of your report and on your ability to communicate (i.e. full sentences, proper grammar, well-formed arguments.)
- Submit your R and STAN code in an accompanying file as an appendix to your report. Each line of code should be commented in a way that explains the purpose of each step to your reader.
- All page-lengths are "highly recommended guidelines," but are not absolutely necessary. (Please don't submit something much longer than what we ask for in order to ensure quick turn-around time for grades.)

Introduction: (≈ 2 pages)

- Describe the general idea behind a meta-analysis and why it is used.
- Briefly describe the intuition behind why the following factors must be accounted for in a meta-analysis.
  - The number of trials included in the meta-analysis.
  - Which trials are included, and how the quality of the trial may impact our beliefs in the results of the meta-analysis.

- Why it is important to account for varying sample sizes across trials (i.e. why results from very large trials should potentially have more of an impact in the meta-analysis than smaller trials)
- How does publication bias impact the selection of studies for a meta-analysis?

- Summarize the differences between Bayesian and frequentist views of statistics.

- Why does the problem of meta-analysis lend itself well to the Bayesian paradigm?

- In one paragraph, briefly introduce the controversy surrounding the use of intravenous magnesium in patients with acute myocardial infarction.

Introduce the models: (≈ 2 - 3 pages)

- Describe the philosophical difference between fixed effects and random effects meta-analysis. How do these methods differ in how they treat variability in the outcomes of different trials? What are some of the potential trade-offs between these two methods? You may find this video helpful: https://www.youtube.com/watch?v=Vb0GvznHf8U

- How does the sample size of each trial impact its weight in the estimation of the overall meta-analytic treatment effect?

- Look at the dataset provided. Describe the studies included in this dataset. Note that the last trial (ISIS-4) is much larger than any of the other trials. How may the role of inclusion of the ISIS-4 mega-trial impact the results of the meta-analysis?

- Describe the model specification for the Peto method for (frequentist) fixed effects meta analysis. Your descriptions should include technical (mathematical) details of the model specification.

- Describe the model specification for the DerSimonian and Laird (D-L) method for (frequentist) random effects meta analysis. Your descriptions should include technical (mathematical) details of the model specification.

The Frequentist analysis: (≈ 2 pages including tables and figures)

- Reproduce table 2 from Higgins and Spiegelhalter paper. You may use the `rma.peto` function in library `metafor` and the `meta.DSL` function in the library `rmeta`. (Don't worry about reproducing the p_{het} value.)

- Interpret the results you obtained from each model. What does an odds ratio of one mean? What does an odds ratio less than one mean? greater than one?

- Under both the fixed effects and random effects models, how do the conclusions you obtain change when subsequent trials are added? How do the results from each method compare to each other? Which model do you think is preferable?

- For each of the three fixed effects models you fit, use the `forest` function in library `metafor` to produce a forest plot. Comment on how the addition of more trials impacts your estimate of the odds ratio.

The Bayesian analysis: ($\approx$ 3 pages including tables and figures)

- Adapt the code from the end of the Higgins and Spiegelhalter paper to run two Bayesian random effects meta analysis models in STAN. One should use the reference prior and one should use the skeptical prior. Make sure to include the specification of your models in mathematical form in your write-up. Note that the code at the end of the paper was written for WinBUGS, and that while WinBUGS uses a $N(\mu, \tau^2)$ specification of a normal distribution (where $\tau^2 = 1/\sigma^2$ is the precision of the distribution), STAN uses a $N(\mu, \sigma)$ specification of a normal distribution. In each model, save only the simulations for the parameter δ_{new}, the predicted values for the overall, meta-analytic, log odds ratio.

- For each choice of prior distribution, use 3 MCMC chains, with at least 500,000 iterations on each chain (unless your computer can't handle this, then run fewer), with an appropriate warm-up period for each chain. Check the convergence of your MCMC chains by visually examining the trace plots. Include the trace plots for δ_{new} under both priors in your write-up.

- For each of the Bayesian models (that is, the two choices of priors), plot a histogram of the posterior distribution for the overall meta-analytic estimate of the odds ratio. Make sure to adjust your `xlim` so that the histogram shows only the meaningful part of the axis. Interpret the appropriate summary statistics from your posterior distribution in the context of the scientific problem.

- Use the MCMC simulations to answer the question: under each of the two prior distributions what is the posterior probability that the overall estimate of the odds ratio is smaller 1 (statistical superiority)? Under each of the two prior distributions, what is the posterior probability that the overall estimate of the odds ratio is smaller than 0.9 (clinical superiority)?

Conclusion: ($\approx$ 1 -2 pages)

- Summarize the methods used in this project. How do their results differ? Which do you think is the most appropriate for this analysis?

- What do you believe is the overall answer to the question of interest? Is magnesium effective in preventing acute myocardial infarction?