MAS8384 Bayesian Methodology Project

Clement Lee
clement.lee@newcastle.ac.uk

Semester 2, 2023/24

Submission

Upload your submission to Canvas by **2024-02-23 Friday 4pm**. You should include the following:

1. the report in pdf format, with a page limit of 12 pages; this should include written work and any plots you have produced;
2. the script file in Rmd format; your JAGS model code and R code are to be included in this file;
3. a short video, approximately 3 minutes long, in which you will discuss one model from your project; this video will be pass/fail and otherwise will not count towards your final mark for the project.

These three files should **not** be compressed in a zip folder. If you have a whole set of supplementary files e.g. if you are using `ProjectTemplate`, you can compress and upload them in a zip folder.

Data description

The project is on some data on cross country race times for male athletes in the North East Harrier League (NEHL). The website for the NEHL, where the data have been taken from, is http://www.harrierleague.com.

The data provide the race times (in minutes) for runners in NEHL races in the years 2016-18. In total there were 17 races over this period.

The data contains the following variables.

- **Number**: Athlete identifier.
- **Age**: The age groups of the athlete: Under 20, Senior, Veteran over 35, Veteran over 40, etc.
- **Pack**: The pack of the athlete: slow (S), medium (M) or fast (F). All athletes begin in the slow pack. If they finish in the top 10% of the field they are promoted to the medium pack for the rest of the year. If they finish in the top 10% of the field from the medium pack they are promoted to the fast pack for the rest of the year.
- **Course**: The location of the race.
- **Year**: The year of the race.
- **Temperature**: The temperature of the race in degrees celcius.
- **Windspeed**: The windspeed during the race in miles per hour.
- **Distance**: The distance of the race in miles.
- **Elevation**: The total metres of ascent during the race.
- **Response**: The time taken in minutes to complete the race.

The data contains almost 8000 rows. This will result in rather slow MCMC chains, and so you will consider only a subset of 1000 rows from the data. To do so, run the following commands in R.

```r
set.seed(n) # replace n by your student ID
run <- read.table("rundata.txt", header = TRUE)
run <- run[sample(1:nrow(run), 1000), ]
```

The value for n you should use in your code is your student ID, which should be a 9-digit number. (It is not the ID that starts with "c", which is your login ID).

Question

Perform a Bayesian analysis of the NEHL dataset, in particular how the different courses appears to affect the time taken to complete the races. The project should be written up as a coherent report on this problem. You will be marked for how well you apply Bayesian methods, including interpretation of their results.

Your report should include:

- Description of various models you consider, your the final model, and why it was selected.
- The diagnostics you use.
- Summaries of your posterior distributions for various models.
- Conclusion. Give a brief non-technical summary of your findings.

This dataset is complicated enough that there is no single best model. There are many different ideas from the course which can try including in your model. An inital model could be to include only random effects for the athlete and the course. You will receive credit for considering sensible models beyond this initial regression, even if they do not turn out to fit well.