University of Edinburgh
School of Mathematics
Bayesian Data Analysis, 2019/2020, Semester 2
Daniel Paulin & Nicolò Margaritella

March 23, 2020

Assignment 2

- To be uploaded to Learn by 17:00, MONDAY 13 April, 2020.
- This assignment is worth 50% of your final grade for the course.
- Assignments should be typed (LaTeX, Word, etc.) and should be no more than 10 pages using a type size not smaller than 11 point and with 1.5-2.0 line space. This includes figures but excludes the appended code. Document your code so that someone can read it without too much guesswork.
- Answers to questions should be in full sentences.
- Any output (e.g., graphs, tables) from `R/JAGS` that you use to answer questions must be included with the assignment. You will want to be judicious with what you include in the written report—not every figure and table you construct needs to be included. Also, please append your `R/JAGS` code at the end of the assignment.
- The assignment is out of 100 marks.
- The main topics of this assignment are Bayesian GLM and Bayesian hierarchical models which are presented in Lecture 4 and 5.
- You are expected to work independently and not to discuss the assignment with others.
- Briefly indicate the technical details of the MCMC analyses you perform (number of iterations, convergence checks, etc.)

1. *Modelling the number of deaths by avalache in the Italian Piedmont region* **(49 marks)**

   The file `Avalanches.csv` contains data regarding the number of avalanches and the relative number of casualties in the Italian Piedmont region during the period 1985-2019. In the first instance, we are interested in the identification and analysis of patterns which could help understand how the number of deaths evolved in time and if there are factors which might have affected such development.

   To this regard, in 1994, the European Avalanche Warning Services introduced a uniform scale called European Avalanche Danger Scale (EADS). Until then, countries had used different scales with varying danger ratings. The adoption of a uniform European Avalanche Danger Scale was considered extremely beneficial for all snow users, professional and amateur, because they could refer to the same danger ratings when visiting other countries. From 2004, daily measurements of the EADS scale have become available online on various Italian governative websites and easily accessible to the general public.

```r
# Loading the data
Avalanches<-read.csv("Avalanches.csv",header = TRUE, sep=";",fileEncoding="UTF-8-BOM")
```

(a) **(5 marks)** Add to the dataset two dummy variables called EADS1 and EADS2 which take both values zero before 1994. EADS1 should take value 1 between 1994 and 2003 and 0 otherwise while EADS2 should take value 1 from 2004 and 0 otherwise. Print the dataset rows relative to the years (`Season`) 1986, 1994, 2004. Perform some exploratory analysis including a graph showing the temporal evolution of the number of avalanches (reported events, `Rep.events`) and deaths (`Deaths`) in the periods considered (1986-1993, 1994-2003, 2004-2019). Does the association of these variables change between those three periods? Briefly comment your results.

(b) **(10 marks)** Fit a Bayesian Poisson model with logarithm link function where the number of deaths (`Deaths`) is regressed on the number of avalanches (`Rep.events`) and the indicator variables `EADS1` and `EADS2`. Use wide normal priors for all the parameters and briefly discuss the results of your convergence checks. Report and interpret your posterior estimates (the file `Q&A Lecture 4_UPDATED.pdf` on Learn can help interpret your results).

NOTE: throughout the assignment you are not supposed to center indicator variables so that you can keep a straightforward interpretation of all your parameters. However, you should center all continuous covariates!

(c) **(9 marks)** Use the output from model 1b to answer the following questions:

i. What is the probability of observing less than 15 deaths if the number of avalanches in Piedmont during 2020 totals 20?

ii. What is the probability of observing more than 1 death per event before the introduction of the EADS scale? Does it change after the EADS introduction? And after its daily publication online?

(d) **(5 marks)** Suppose we are told by experts that the number of avalanches per year in Piedmont (`Rep.events`) is usually between 5 and 15. In addition, for an extreme number of events, they consider it plausible that the number of deaths could be 4 times greater/smaller than the usual (average) number of casualties. Use this information to check if a logNormal prior distribuion with mean 0 and SD=2 for $\phi = \exp((x - \mu_x) * \beta_{Rep.events})$ is a sensible choice and work out if the relative induced prior for $\beta_{Rep.events}$ can also be appropriate (show your computations).

(e) **(10 marks)** Expand the previous model in (1b) by including an extra variance term theta in the covariates to capture the high variability in the number of deaths for different years. Set a Gaussian prior for theta with mean 0 and standard deviation with a uniform hyperprior distribution between 0 and 10. Compare your results with those in (1b).

(f) **(10 marks)** For the model in (1b) and (1e), compute the posterior distribution of the number of deaths for each value of the covariates in the dataset. Plot the posterior means and 90% point-wise credible intervals into two separate graphs (you could use `par(mfrow=c(2,1))` command in `R`). Compute the Deviance Information Criterion

(DIC) of the two models (hint: look at tutorial 4 to find out how this is done in JAGS). By looking at the new graphs, the plot in (1a), and the DIC, comment the performance of the two models and indicate which you would prefer and why.

2. *Modelling the probability of a deadly avalanche with snow abundance and permanence from several recording locations in the Piedmont mountains.* **(51 marks)**

   The second dataset we consider is `Avalanches_part2.csv`. For each avalanche in the Piedmont region between 2014 and 2019 (`Season`), it reports the number of people involved (`Hit`) and killed (`Deaths`) together with the location of the nearest recording station (`Rec.station`). At each recording station, the total amount of snow (`Snow_total`, cm) and its permanence (`Snow_days`, days) are recorded for each year. The location of the 11 recording stations divided into three geographical areas (`Geo.space`) can be seen in Figure 1.

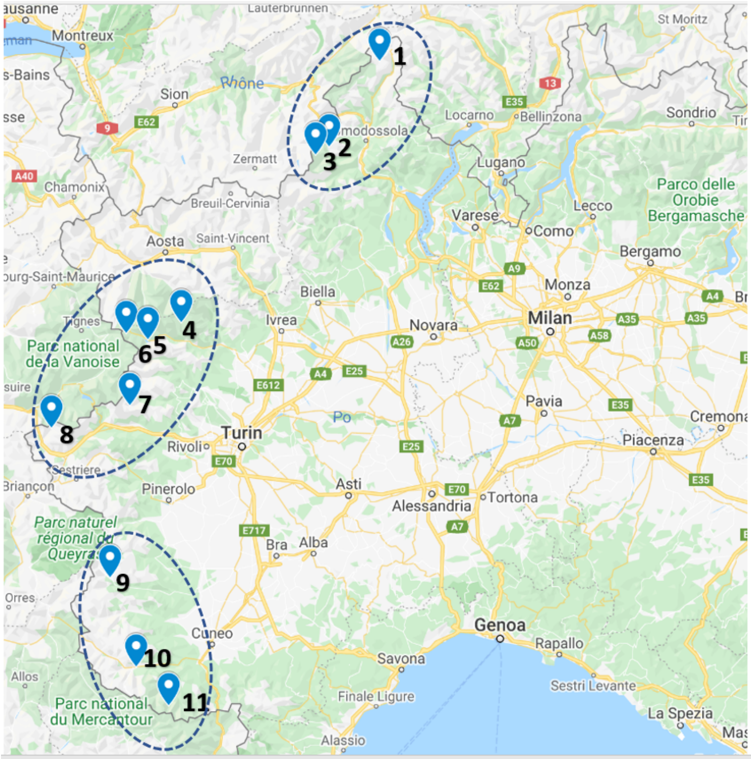

Figure 1: Snow recording stations in the Italian Piedmont region are divided into three mountain areas: north, centre and south.

In this analysis we want to explore whether the proportion of people killed by an avalanche in Piedmont can be associated to the amount of snow and its seasonal permanence on the relative mountain area.

```r
# Loading the data
Avalanches_part2<-read.csv("Avalanches_part2.csv",
                  header = TRUE, sep=";",fileEncoding="UTF-8-BOM")
```

(a) **(4 marks)** Transform the variables `Snow_total` and `Snow_days` into meters and fortnights, respectively. Print the head of your transformed dataset. Perform some exploratory analysis of the database, including an analysis of the association between the variables `Season`, `Snow_total` and `Snow_days`. Briefly comment your results.

(b) **(15 marks)** Fit a Bayesian hierarchical binomial logistic model where the proportion of deaths among those hit by an avalanche is explained by `Season`, `Snow_total` and `Snow_days` and a random effect on the mountain area (`Geo_space`). Use normal prior for all covariates with variance = 10. Use a Gaussian prior for the random effects with mean zero and standard deviation with a uniform hyperprior distribution between 0 and 10. Write the hierarchical formula for this model and the DAG (a well-drawn sketch). Discuss the posterior estimates obtained.

(c) **(6 marks)** Run the model again without `Snow_Days` and compare the results of both diagnostic checks and posterior estimates with those in 2b.

(d) **(12 marks)** Based on the model in 2c, estimate the posterior expected value and 95% credible interval of the proportion of deaths near recording stations 1, 8 and 10 for the Seasons 2015 and 2018. Compare the probability of a proportion of deaths greater than 60% between the three stations in the two years considered. Interpret your findings.

(e) **(10 marks)** Set up a new hierarchical model where the random effects are placed on the recording stations (`Rec.station`). Use the same prior and hyperprior distributions as in 2c and compare the results using DIC as well. As in 2d, estimate the posterior expected value and 95% credible interval of the proportion of deaths near recording stations 1, 8 and 10 for the Seasons 2015 and 2018. Briefly discuss your results.

(f) **(4 marks)** Could we capture variability in the different recording stations and mountain areas in one single hierarchical model? Write down your proposed formula and the relative DAG (if you feel brave, you could try to code it!)