BAYESIAN STATISTICAL MODELS - University of Milano-Bicocca

Case study - the frogs data

Lecturer: Riccardo Corradin

Case study 1
We consider a set of data contained in the file *frogs.csv*, available at the module page. The data consist in a set of measures for different amphibious. Specifically, for each observed animal we measure

- The scientific *family*.
- The *body weight*.
- The *nose-to-tail length*.
- The *brain weight*.

We are interested study the body weight, here playing the role of the response variable, as function of the other observed quantities. Specifically, we first transform on suitable scales the positive real-valued observed quantities, then we consider a linear model of the form

$$y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4},$$

where $x_{i,1} = 1$ for all $i = 1, \dots, n$, $x_{i,2}$ is a binary variable denoting if an observation is in the *Ranidae* family or not, $x_{i,3}$ is the logarithm of the nose-to-tail length, $x_{i,4}$ is the logarithm brain weight. We assume an hierarchical model of the form

$$\begin{aligned} Y_i \mid \boldsymbol{x}_i, \boldsymbol{\beta}, \sigma^2 &\sim N(\boldsymbol{x}_i^\intercal \boldsymbol{\beta}, \sigma^2), \qquad i = 1, \dots, n, \\ \boldsymbol{\beta} \mid \sigma^2 &\sim N(\boldsymbol{b}_0, \sigma^2 \Sigma_0), \\ \sigma^2 &\sim IG(a_0, b_0). \end{aligned}$$

a) Implement a function in R to sample from the posterior distribution of interest.

b) Produce a sample of size $1\,000$ from the posterior distribution of interest.

c) For each regression coefficient, obtain point estimate and credible interval of your choice, and compute
$$\min\{P(\beta_j > 0), P(\beta_j < 0)\}.$$
Use the previous to identify which coefficients have no significative effect in the model. You can obtain the credible intervals, for example, with the *ci* function in *bayestestR* library.

d) Produce a second model estimate without the covariates previously identified, check the posterior distributions of interest. Then, perform a test to compare the two models.

Case study 2

Consider a set of 403 observations, where each observation correspond to a particular grouse. You can find the data in the file *grouses.csv*, available at the module page. Specifically, for each observed animal we measure

- Height, above sea level (meters), denoted by X , where the animal lives, continuous variable.
- Year (three different years), denoted by G , where G = 1 is 1994, G = 2 is 1995 and G = 3 is 1996.
- Number of ticks Y on the heads of the red grouse, which plays the role of response variable.

We are interested study the number of ticks on the grouses' head, as function of the other observed quantities. Specifically, we want to consider a Poisson regression model including a dummy variable specific for each year. The model specification is then given by

$$
\begin{aligned}
Y_i \mid \mu_i &\sim Poi(\mu_i), \qquad i = 1, \dots, n \\
\mu_i &= \mathrm{e}^{\beta_1 + \beta_2 x_i + \beta_3 \mathbf{1}_{[G=2]} + \beta_4 \mathbf{1}_{[G=3]}}, \qquad i = 1, \dots, n, \\
\boldsymbol{\beta} &\sim N(\mathbf{0}, \Sigma_0),
\end{aligned}
$$

where Y_i denotes the ith observation, μ_i the ith expectation, and $(\beta_1, \beta_2, \beta_3, \beta_4)$ are our regression coefficients.

a) Implement the model in STAN. Think wisely about how to include the random effects.

b) Choose specific values for the dispersion parameters of the prior distributions, Σ_0 and σ_0^2, and produce a sample of size $5\,000$ from the posterior distribution of interest, of which $2\,000$ are burnin values.

c) Check if the random coefficients are significantly different from 0, by looking at their credible intervals.

d) Produce a plot that shows the model behavior for each separated year. Ideally, you should show a plot with three line, one for each year, that corresponds to the evolution of the Poisson distribution expectation as far as the covariate is moving on its support. Hint: restrict the values you are considering of the covariates on the observed range.

e) Produce a second model estimate without the random effects. Test if the two models are significantly different from each other.