BAYESIAN STATISTICAL MODELS - University of Milano-Bicocca

Case study - improvement

Lecturer: Riccardo Corradin

We consider the dataset *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. We first transform on suitable scales the positive real-valued observed quantities, then we consider a linear model with the informative prior assumptions, i.e.

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

1) Produce a code that estimate all the possible models, computing also the WAIC and LPML indices.
2) Test the best model you identify against the saturated model and the model including only the intercept term.
3) Compute the Bayesian model average estimate for the regression coefficients.
3) Compute the variance of the BMA estimate for the regression coefficients, marginally for each coefficient.

We consider a set of $n = 681$ observations from the Rotterdam Early Arthritis Cohort (REACH) dataset. The study was initiated in 2004 to investigate the development of rheumatoid arthritis in patients with early manifestations of joint impairment. Information regarding basic patient characteristics, serological measurements, and patterns of disease involvement at baseline has been gathered for the 681 recruited patients. It is of interest to know which of the following 12 factors are potentially associated with the development of rheumatoid arthritis considered as a binary (yes/no) outcome.

- ACCP (cyclic citrullinated peptide antibody)
- age
- ESR (erythrocyte sedimentation rate, is the rate at which red blood cells sediment in a period of one hour)
- DC (duration of complaints in days)
- stiffness (duration of morning stiffness in minutes)
- RF (rheumatoid factor)
- gender
- Sym (symmetrical pattern of joint inflammation; yes/no)
- SJC (swollen joint count)
- TJC (tender joint count)
- BCPH (bilateral compression pain in hands; yes/no)
- BCPF (bilateral compression pain in feet; yes/no)

We consider a probit model of the form

$$\begin{aligned} Y_i \mid \theta_i &\sim Be(\theta_i), \qquad i = 1, \ldots, n, \\ \theta_i &= \Phi(\beta_0 + \boldsymbol{x}_i^\intercal \boldsymbol{\beta}), \\ \beta_0 &\sim N(0, \tau^2), \qquad \beta_j \sim \pi N(0, \tau^2) + (1-\pi) N(0, c\tau^2),\ j = 1, \ldots 12, \\ \pi &\sim Beta(\alpha_1, \alpha_2), \end{aligned}$$

where $\tau^2 = 10^3$, $c = 0.001$, $\alpha_1 = 1$ and $\alpha_2 = 1$.

1) Implement the model in STAN.
2) Produce a sample of $10\,000$ realizations from the posterior distribution of interest. Check the diagnostic of the model.
3) Find the best model according to HPD, MPM and HS criteria.

Now, we consider a set of 60 observations of an ecological experiment. Each observation corresponds to a specific tree. The trees are divided into High and Low altitude trees, defining two macro-groups. The experiment wants to study the amount of ants that are populating each specific tree. For each spot we observe:

- Trap days, different lengths of time surveying each tree, denoted by X.
- Elevation, G . If G = 1, then the spot is High, if G = 2 then the spot is Low.
- Valley, three different valleys for each elevation, denoted by SG, nested within each elevation.
- Summer precipitation (avg), assuming a random effect, denoted by Z1.
- Summer temperature (avg), assuming a random effect, denoted by Z2.
- Number of ants,denoted by Y.

We assume a priori a Poisson distribution of the data, having a model of the form

$$
\begin{aligned}
Y_{i,j,r} \mid \mu_{i,j,r} &\sim Poi(\mu_{i,j,r}), \qquad i = 1, \dots, n_j,\ j = 1, 2,\ r = 1, 2, 3, \\
\mu_{i,j,r} &= \exp\{\xi_0 + \beta x_{i,j,r} + \theta_j + \gamma_{j,1} z_{1,i,j,r} + \gamma_{j,2} z_{2,i,j,r} + \alpha_{j,r}\}, \qquad i, j, r = \dots, \\
\xi_0 &\sim N(0, \psi_0^2), \\
\beta &\sim N(0, \sigma_0^2), \\
\theta_j &\sim N(0, \eta_0^2), \qquad j = 1, 2, \\
\gamma_{j,1}, \gamma_{j,2} &\sim N(0, \tau_0^2), \qquad j = 1, 2, \\
\alpha_{j,r} &\sim N(0, \lambda_0^2)
\end{aligned}
$$

where β is the fixed effect, associated to the trap days, θ_js are random intercepts specific for each elevation, $\alpha_{j,r}$ are nested random intercepts specific for each valley, $\gamma_{j,1}$ and $\gamma_{j,2}$ are random effects associated to summer precipitation and temperature.

1) Implement the model in STAN.
2) Choose specific values for the dispersion parameters of the prior distributions, σ_0^2, η_0^2, τ_0^2, λ_0^2. Produce a sample of size $5\,000$ from the posterior distribution of interest, of which $2\,000$ are burnin values. Looking at the traceplots, think about the model specification and possibly discard some random intercepts.
3) Check if the random intercepts, θ_js and $\alpha_{j,r}$s, are significantly different from 0, by looking at their credible intervals.
4) Show the posterior distribution of the estimated coefficients with a suitable plot.
5) Produce a second model estimate without the random effects, only having a common intercept term and fixed effects for X, Z1 and Z2. Test if the two models are significantly different from each other.