Bayesian Data Analysis

Exercise 2

Stack loss data.

21 daily responses of stack loss, y_i, the amount of ammonia escaping from industrial chimneys.
Covariates: air flow x_1, temperature x_2 and acid concentration x_3.
The question of interest is to identify the factors that affect the amount of ammonia escaping from industrial chimneys.
Statistical analysis.
Transform the covariates, so that they have the same sample mean and sample variance:

$$z_{ki} = (x_{ki} - \bar{x}_k)/sd(x_k), \quad k = 1, 2, 3.$$

Possible likelihood in terms of the transformed covariates:

$$y_i \mid \beta_0, \beta_1, \beta_2, \beta_3, \sigma \sim N(\beta_0 + \sum_{k=1}^{3} \beta_k z_{ki}, \sigma^2), i = 1, 2, \ldots, n \quad \textit{independently.}$$

The data, some initial values and a basic model are provided in stacks-data.odc, stacks-model.odc, stacks-init1.odc and stacks-init2.odc, respectively.

1. State how you can address the question of interest, to identify the factors that affect the amount of ammonia escaping from industrial chimneys, in terms of β_k.

2. Consider proper 'non-informative' independent priors for parameters β_k, $k = 0, \ldots, 3$: $\beta_k \sim N(0, A^2)$, $k = 0, 1, 2, 3$ independently, with $A = 1000$. Can you justify why these priors can be referred to as 'non-informative'?

3. **Priors for σ.**

   (a) Consider a proper 'non-informative' prior for precision $\tau = \sigma^{-2}$: $\tau \sim \Gamma(\epsilon, \epsilon)$ with $\epsilon = 0.001$. Discuss why this prior can be viewed as a 'non-informative', and other proper 'non-informative' priors for σ.

   (b) Use a priori information from a previous study that the average precision is known to be 0.12, to specify a prior for σ. The variability is not known a priori so try two different values of the prior variance (one with a large variance, e.g. 10, and another with a smaller variance, e.g. 1). [*Hint: typical prior distributions for precision $\tau = \sigma^{-2}$ are Gamma and log Normal distributions. Don't forget to state the expressions for the mean and for the variance of these distributions.*]

4. **Posterior analysis.** For each of the considered priors for σ, run a WinBUGS model (using at least two chains and running the chains until convergence):

   (a) produce the posterior summaries (mean, median, standard deviation, and two-sided 95% credible interval) for β_k, $k = 0, \ldots, 3$ and σ;

   (b) produce the density plots of the posterior distribution of β_k, $k = 0, \ldots, 3$ and σ;

(c) use the posterior distribution of β_k, $k = 0, \ldots, 3$ to address the question of interest (give the conclusion in terms of the original question).

5. **Convergence.** For each prior for σ,

   (a) Produce the history plots for parameters β_k, $k = 0, \ldots, 3$ and σ. Justify the argument that the MCMC converged at the burnin iteration (and hence your use of the number of burnin iterations).

   (b) Produce the plot of the Gelman-Rubin statistic for β_k, $k = 0, \ldots, 3$ and σ. What conclusion can you make about convergence of the MCMC?.

   (c) Produce the autocorrelation plots for parameters β_k, $k = 0, \ldots, 3$ and σ. Comment on the strength of the correlation (and whether it is necessary to thin.)

   (d) Produce the autocorrelation plots when fitting the model to non-standardised covariates, i.e. for the following likelihood:

   $$y_i \mid \beta_0, \beta_1, \beta_2, \beta_3, \sigma \sim N(\beta_0 + \sum_{k=1}^{3} \beta_k x_{ki}, \sigma^2), i = 1, 2, \ldots, n \quad \textit{independently.}$$

   How do the plots change?

   (e) Give the MC error for parameters β_k, $k = 0, \ldots, 3$ and σ. Comment how the MC error compares with the standard deviation of the posterior distribution. (You can give the MC error in the posterior summary table in 4(a), and comment on the MC error here.)

6. **Sensitivity to the prior.**

   (a) Compare the posterior distributions of the parameters for different priors, and comment whether they are affected by the different choices of prior.

   (b) Comment whether the main conclusion on the question of interest (4(c)) is affected by the choice of priors.

7. **Model checks.** Residual plots (Pearson residuals).

   (a) Add in the code to calculate the standardised residuals $r_i = (y_i - \mu_i)/\sigma$ where $\mu_i = \beta_0 + \sum_{k=1}^{3} \beta_k z_{ki}$;

   (b) produce the residual plots; see section Plotting summaries of the posterior distribution - Box and caterpillar plots of the hints handout.

   (c) Discuss the residuals, and whether there is any evidence of the outliers. Use t_4 distribution as the distribution of errors instead of the normal distribution, and produce the residual plots as in 7(b) and (c). Is there any improvement?