Take-home midterm Exam. Due Date: Monday, 28th March

Please submit all your codes for this Midterm by e-mail to the TA Freda Cooner at xiyunwu@biostat.umn.edu

1. Consider the data set given on the web at http://www.biostat.umn.edu/~sudiptob/pubh5485/LinearModelExample.txt. Concentrate only on the response and the \( N \times p \) covariance matrix, \( X \), formed by the intercept and six covariates: \( X_1 \) through \( X_6 \). Thus, \( p = 7 \). Using the completely non-informative prior, \( p(\beta, \sigma^2) \propto 1/\sigma^2 \). Write R programs to fit models by sequentially adding covariates: (1) Intercept only (2) Intercept + \( X_1 \) (3) Intercept + \( X_1 + X_2 \) (4) Intercept + \( X_1 + X_2 + X_3 \) (5) Intercept + \( X_1 + \ldots + X_4 \) (6) Intercept + \( X_1 + \ldots + X_5 \) (7) Intercept + \( X_1 + \ldots + X_6 \). Present the 95% posterior credible interval for the coefficients for the slopes in each of the above models. Which of the six slopes (other than the intercept) is the most significant?

For each of the above models, suppose you have generated posterior samples of size \( M \), say \((\beta_l, \sigma^2_l)\), \( l = 1, \ldots, M \). Based upon these samples generate \( M \) replicated data sets \( y^l_{\text{rep}}, l = 1, \ldots, M \). Now consider the omnibus statistic:

\[
T(y, \beta, \sigma^2) = \sum_{i=1}^{N} \frac{(y_i - x_i \beta)^2}{\sigma^2},
\]

where \( N \) is the number of observations, and \( x_i \) is the \( i^{th} \) row of the covariance matrix \( X \). Write an R program to compute \( T(y^l_{\text{rep}}, \beta_l, \sigma^2_l) \) and \( T(y_{\text{obs}}, \beta_l, \sigma^2_l) \) for \( l = 1, 2, \ldots, M \). Calculate the following quantity:

\[
\frac{1}{M} \sum_{l=1}^{M} 1(T(y^l_{\text{rep}}, \beta_l, \sigma^2_l) > T(y_{\text{obs}}, \beta_l, \sigma^2_l)),
\]

where \( 1(A) \) is the indicator function taking the value 1 if the event is true, and 0 if it is not true. Note You have just calculated a Monte Carlo estimate of the Bayesian \( p - \text{value} \) defined by:

\[
P(T(y_{\text{rep}}, \theta) > T(y_{\text{obs}}, \theta) | y_{\text{obs}}).\]

Present the Bayesian p-values for each of the models (1)–(7) and order them in terms of closeness to 0.5. Can you comment on the consistency between this p-value criteria and the significance of the covariates in as far as proposing an optimal set of covariates is concerned?
2. Consider the data set given on the web at http://www.biostat.umn.edu/~sudiptob/pubh5485/BinaryRegressionExample.txt. This is an ecological data set, where the presence or absence of a species is recorded as a binary variable Copresence (1 for presence, 0 for absence). Potentially interesting covariates include CanopyClosure (ordinal variable 0–3), Heavily ManagedPts (binary variable 0–1) and LogEdgeDistance (continuous variable, units- meters). You may ignore all the other variables in the data set. Write a WinBUGS program to model \( P(\text{Copresence} = 1) \) as a logistic regression:

\[
\log \left( \frac{P(\text{Copresence} = 1)}{1 - P(\text{Copresence} = 1)} \right) = \beta_0 + X_1 \beta_1 + X_2 \beta_2 + X_3 \beta_3,
\]

where \( X_1, X_2 \) and \( X_3 \) are CanopyClosure, HeavilyManagedPts and LogEdgeDistance respectively. Present your results in terms of the credible intervals for the parameters.

Next, modify your program to incorporate random effects for each location that are identically and independently distributed as \( N(0, \sigma^2) \) in the logistic regression. What would be a good way to present the posterior estimates of the random effects? Do you think random effects are needed in explaining the variation in Copresence over and above the existing covariates?