BIOS 779

Spring, 2023

Homework 5

Ibrahim

1. Derive all of the full conditional distributions for β, σ^2, and γ for the variable subset selection method of George and McCulloch (1993) given on pages 544-551 of the notes.

2. Consider a *general* regression model, and let $p(y_i|\theta, x_i)$ denote the sampling density of y_i for case i and x_i is the $p \times 1$ vector of covariates for subject i. Assume that the observations y_i (given theta) are independent for $i = 1, \ldots, n$. Let $p(\theta|y, X)$ denote the posterior density of θ from this model, where θ is $p \times 1$, X is $n \times p$, and y is $n \times 1$. Show that
$$CPO_i = \left\{ E_{\theta|y,X} \left(\frac{1}{p(y_i|\theta, x_i)} \right) \right\}^{-1}.$$

3. We consider data on $n = 136$ patients from a liver cancer clinical trial. Here, we are primarily interested in the patient's status as he/she enters the trial. In particular, we are interested in how the number of cancerous liver nodes (y) when entering the trial is predicted by six other baseline characteristics: body mass index (x_1), (defined as weight in kilograms divided by the square of height in meters), age in years (x_2), time since diagnosis of the disease in weeks (x_3), two biochemical markers (each classified as normal=1 or abnormal=0): Alpha fetoprotein (x_4), and Anti Hepatitis B antigen (x_5), and associated jaundice (yes=1, no=0) (x_6). Let x_{ij} = the j^{th} covariate measurement on the i^{th} subject, $i = 1, \ldots, n$, $j = 1, 2, 3, 4, 5, 6$. In all of the computations below, standardize all covariates by
$$z_{ij} = (x_{ij} - \bar{x}_j)/sd_j$$
for $j = 1, \ldots, 6$, and $i = 1, \ldots, n$, where $sd_j = \left(\frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}{n-1} \right)^{1/2}$.

   We consider a Poisson regression model (with a canonical link) for the data, with an intercept and the covariates $(z_1, z_2, z_3, z_4, z_5, z_6)$, and thus our regression coefficient vector is $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6)$.

   (a) Write out the likelihood function of β for this model.

   (b) Compute the maximum likelihood estimate of β and the standard errors of the estimates. Note: You can do this in R.

   (c) Consider the conjugate prior for β with $a_0 = 1$ and $y_0 = (1, \ldots, 1)$. Use Stan to obtain samples from the posterior distribution of β. Attach all of the Stan code for doing all computations. Using your posterior samples from Stan, compute

i) the posterior mean and covariance matrix of β.

ii) the 2.5%, 50%, and 97.5% posterior percentiles of $(\beta_0, \beta_1, \ldots, \beta_6)$.

iii) Use Stan to plot the marginal posterior distributions of β.

iv) Use Stan to assess convergence diagnostics.

(d) Repeat part (c) using a uniform improper prior for β. Run both Stan and the Bayes procedure in SAS GENMOD to summarize the results.

(e) Consider a normal prior $\beta \sim N(0, (X'X)^{-1})$. Repeat part (c) using SAS GENMOD, and compare your answers in parts (c), (d), and (e).

(f) Using the prior in c), use the Bayesian central limit theorem to derive the asymptotic posterior distribution of β.

(g) Carry out a variable subset selection procedure of these data using R or SAS. Specifically, identify the top 10 models based on both AIC and BIC. For each of these models, compute the DIC, the L measure, and $B = \sum_{i=1}^{n} \log(CPO_i)$. For DIC, the L measure, and the B statistic, assume a uniform improper prior for β.