BIOSTAT/STAT 570: Coursework 5

To be submitted to the course canvas site by 11:59pm Monday 4th November, 2022.

1. Consider the data given in Table 1, which are a simplified version of those reported in Breslow and Day (1980). These data arose from a case-control study that was carried out to investigate the relationship between esophageal cancer and various risk factors. Disease status is denoted Y with $Y = 0/1$ corresponding to without/with disease and alcohol consumption is represented by X with $X = 0/1$ denoting $< 80g/ \geq$ 80g on average per day. Let the probabilities of high alcohol consumption in the cases and controls be denoted

$$p_1 = \Pr(X = 1 \mid Y = 1) \quad \text{and} \quad p_2 = \Pr(X = 1 \mid Y = 0),$$

respectively. Further, let X_1 be the number exposed from n_1 cases and X_2 be the number exposed from n_2 controls. Suppose $X_i \mid p_i \sim$ Binomial(n_i, p_i) in the case ($i = 1$) and control ($i = 2$) groups.

	$X = 0$	$X = 1$	
$Y = 1$	104	96	200
$Y = 0$	666	109	775

Table 1: Case-control data: $Y = 1$ corresponds to the event of esophageal cancer, and $X = 1$ exposure to greater than 80g of alcohol per day. There are 200 cases and 775 controls.

(a) Of particular interest in studies such as this is the *odds ratio* defined by

$$\theta = \frac{\Pr(Y = 1 \mid X = 1)/\Pr(Y = 0 \mid X = 1)}{\Pr(Y = 1 \mid X = 0)/\Pr(Y = 0 \mid X = 0)}.$$

Show that the odds ratio is equal to

$$\theta = \frac{\Pr(X = 1 \mid Y = 1)/\Pr(X = 0 \mid Y = 1)}{\Pr(X = 1 \mid Y = 0)/\Pr(X = 0 \mid Y = 0)} = \frac{p_1/(1 - p_1)}{p_2/(1 - p_2)}.$$

(b) Obtain the MLE and an asymptotic 90% confidence interval for θ, for the data of Table 1.

(c) We now consider a Bayesian analysis. Assume that the prior distribution for p_i is the beta distribution Be(a, b) for $i = 1, 2$. Show that the posterior distribution $\pi(p_1, p_2 \mid x_1, x_2)$ is given by the product of the beta distributions Be$(a+x_i, b+n_i-x_i)$, $i = 1, 2$.

(d) Consider the case $a = b = 1$. Obtain expressions for the posterior mean, mode and standard deviation. Evaluate these posterior summaries for the data of Table 1. Report 90% posterior credible intervals for p_1 and p_2.

(e) Examine the implied prior distribution for θ and give a 90% prior interval.

(f) Simulate samples $p_1^{(t)}, p_2^{(t)}$, $t = 1, \dots, T = 1000$ from the posterior distributions $p_1 \mid x_1$ and $p_2 \mid x_2$. Form histogram representations of the posterior distributions using these samples and obtain sample-based 90% credible intervals.

(g) Obtain samples from the posterior distribution of $\theta \mid x_1, x_2$ and form the histogram representation of the posterior. Obtain the posterior median and 90% credible interval for $\theta \mid x_1, x_2$ and compare with the likelihood analysis.

(h) Suppose the rate of esophageal cancer is 18 in 100,000. Describe how this information may be used to evaluate

$$q_1 = \Pr(Y = 1 \mid X = 1) \quad \text{and} \quad q_0 = \Pr(Y = 1 \mid X = 0).$$

(i) Suppose that *a priori* you would like to select a Be(a, b) distribution on the rate of esophageal cancer with 5% of the mass less than 16 in 100,000 and 5% of the mass greater than 20 in 100,000. Find a and b to satisfy these requirements, and hence obtain samples from the posteriors for q_1 and q_0.

2. (a) Consider the "likelihood", $\widehat{\theta} \mid \theta \sim$ N(θ, V) and the prior $\theta \sim$ N$(0, W)$ with V and W known. Show that $\theta \mid \widehat{\theta} \sim$ N$(r\widehat{\theta}, rV)$ where $r = W/(V + W)$.

(b) Suppose we wish to compare the models $M_0 : \theta = 0$ versus $M_1 : \theta \neq 0$. Show that the Bayes factor is given by

$$\text{BF} = \frac{p(\widehat{\theta}|M_0)}{p(\widehat{\theta}|M_1)} = \frac{1}{\sqrt{1-r}} \exp\left(-\frac{Z^2}{2} r\right)$$

where $Z = \widehat{\theta}/\sqrt{V}$.

(c) Suppose we have a prior probability $\pi_1 = \Pr(M_1)$ of model M_1 being true. Write down an expression for the posterior probability $\Pr(M_1|\widehat{\theta}_1)$, in terms of the BF.

(d) Now suppose we have summaries from two studies, $\widehat{\theta}_j, V_j$, $j = 1, 2$. Assuming, $\widehat{\theta}_j \mid \theta \sim$ N(θ, V_j) and the prior $\theta \sim$ N$(0, W)$, derive the posterior $p(\theta|\widehat{\theta}_1, \widehat{\theta}_2)$.

(e) Derive the Bayes factor

$$\text{BF} = \frac{p(\widehat{\theta}_1, \widehat{\theta}_2|M_0)}{p(\widehat{\theta}_1, \widehat{\theta}_2|M_1)}$$

again comparing the models $M_0 : \theta = 0$ versus $M_1 : \theta \neq 0$.

We will show these results can be used in the context of a genome-wide association study on Type II diabetes, reported bu Frayling et al. (2007, Science). Two sets of data were independently collected, resulting in two log odds ratios $\widehat{\theta}_j$, $j = 1, 2$, for each SNP. For SNP rs9939609 point estimates of the odds ratio (95% confidence intervals) were 1.27 (1.16, 1.37) and 1.15 (1.09,1.23). Suppose we have a normal prior for the log odds ratio that has a 95% range [log(2/3), log(3/2)].

(f) Find W from this interval, and then calculate the posterior median and 95% intervals for θ based on (i) the first dataset only, (ii) both of the populations.

(g) Calculate the Bayes factor based on the first dataset only, and then based on both datasets.

(h) With a prior of $\pi_1 = 1/5000$, calculate the probabilities, $\Pr(M_1|\widehat{\theta}_1)$ and $\Pr(M_1|\widehat{\theta}_1, \widehat{\theta}_2)$

3. We will carry out a Bayesian analysis of the lung cancer and radon data, that were examined in lectures, using INLA. These data are available on the class website.

The likelihood is

$$Y_i \mid \boldsymbol{\beta} \sim_{ind} \text{Poisson}\,[\; E_i \exp(\beta_0 + \beta_1 x_i) \;]\,,$$

where $\boldsymbol{\beta} = [\beta_0, \beta_1]^{\mathsf{T}}$, Y_i and E_i are observed and expected counts of lung cancer incidence in Minnesota in 1998–2002, and x_i is a measure of residential radon in county i, $i = 1, \ldots, n$.

(a) Analyze these data using the default prior specifications in INLA. Produce figures of the INLA approximations to the marginal distributions of β_0 and β_1, along with the posterior means, posterior standard deviations, and 2.5%, 50%, 97.5% quantiles.

(b) For a more informative prior specification we may reparameterize the model as

$$Y_i \mid \boldsymbol{\theta} \sim_{ind} \text{Poisson}\left(E_i \theta_0 \theta_1^{x_i - \overline{x}}\right),$$

where $\boldsymbol{\theta} = [\theta_0, \theta_1]^{\mathsf{T}}$ where

$$\theta_0 = \mathsf{E}[Y/E \mid x = \overline{x}] = \exp(\beta_0 + \beta_1 \overline{x})$$

is the expected standardized mortality ratio in an area with average radon. The parameter $\theta_1 = \exp(\beta_1)$ is the relative risk associated with a one-unit increase in radon.

For θ_0 we assume a lognormal prior with 2.5% and 97.5% quantiles of 0.67 and 1.5 to give $\mu = 0, \sigma = 0.21$. For θ_1 we again take a lognormal prior and assume the relative risk associated with a one-unit increase in radon is between 0.8 and 1.2 with probability 0.95, to give $\mu = -0.02, \sigma = 0.10$. By converting these into normal priors in INLA, rerun your analysis, and report the same summaries.