Bayesian course - problem set 6 (lecture 7)

Ben Lambert

December 7, 2016

1 A meta-analysis of beta blocker trials

Table 1 shows the results of some of the 22 trials included in a meta-analysis of clinical trial data on the effect of beta-blockers on reducing the risk of myocardial infarction [2]. The file "prob5_betaBlocker.csv" contains the full dataset.

The aim of this meta-analysis is to determine a robust estimate of the effect of beta-blockers by pooling information from a range of previous studies.

Table 1: The data from the original study.

	Mortality: deaths/total	
Study	**Treated**	**Control**
1	3/38	3/39
2	7/114	14/116
3	5/69	11/93
4	102/1533	127/1520
...		
20	32/209	40/218
21	27/391	43/364
22	22/680	39/647

Problem 1.1 *Start by assuming that the numbers of deaths in the control (r_i^c) and treated (r_i^t) groups for each trial are given by binomial distributions of the form:*

$$r_i^c \sim binomial(p_i^c, n_i^c) \tag{1}$$

$$r_i^t \sim binomial(p_i^t, n_i^t) \tag{2}$$

where (n_i^t, n_i^c) are the numbers of individuals in the treatment and control datasets respectively. Further assume that the probabilities of mortality in the treatment and control datasets are given by:

$$logit(p_i^c) = \mu_i \tag{3}$$
$$logit(p_i^t) = \mu_i + \delta_i \tag{4}$$
$$\tag{5}$$

where $logit(x) = log\left(\frac{x}{1-x}\right)$*, and we expect* $\delta_i < 0$ *if the beta-blockers have the desired effect. We assume the following diffuse priors for the parameters*

$$\mu_i \sim N(0, 10) \tag{6}$$
$$\delta_i \sim N(0, 10) \tag{7}$$
$$\tag{8}$$

Estimate the posteriors for δ_i *for the above model using Stan, or otherwise. Note: that for this model there is no inter-dependence between the studies. (Hint: use the Stan function "binomial_logit".)*

Problem 1.2 *An alternative framework is a hierarchical model where we assume there to be a common over-arching distribution, across trials such that* $\delta_i \sim N(d, \sigma)$*. By assuming the following priors on these parameters estimate this model:*

$$d \sim N(0, 10) \tag{9}$$
$$\sigma \sim cauchy(0, 2.5),\ for\ \sigma \geq 0 \tag{10}$$

Estimate the posteriors for δ_i *using Stan. How do these estimates compare to the non-hierarchical model?*

Problem 1.3 *Using the hierarchical model estimate the cross-study effect of the beta-blockers. (Hint: use the "generated quantities" code block.)*

Problem 1.4 *For an out of sample trial suppose we know that* $\mu_i = -2.5$*. Using the cross-study estimates for* δ *estimate the reduction in probability for a patient taking the beta-blockers.*

Problem 1.5 *Estimate a model with a single, constant value of* δ *and* μ *across all trials. Graph the posterior for* δ*, and compare it with the cross-study hierarchical model estimate.*

Problem 1.6 *Carry out appropriate posterior predictive checks on the homogeneous and hierarchical models, and hence conclude the preferred modelling choice.*

2 I can't get no sleep

These data are from a study described in Belenky et al. (2003) [1] that measured the effect of sleep deprivation on cognitive performance. There were 18 subjects chosen from a population of interest (lorry drivers) who were restricted to 3 hours of sleep during the trial. On each day of the experiment their reaction time to a visual stimulus was measured. The data for this example are contained within "prob5_sleepstudy.csv", and contains three variables: Reaction, Days and Subject ID which measure the reaction time of a given subject on a particular day.

A simple model that explains the variation in reaction times is a linear regression model of the form:

$$R(t) \sim N(\alpha + \beta t, \sigma) \tag{11}$$

where $R(t)$ is the reaction time on day t of the experiment across all observations.

Problem 2.1 *Assuming normal(0,250) priors on both α and β code up the above model in Stan. Use it to generate 1000 samples per chain, across 4 chains. Has the sampling algorithm converged?*

Problem 2.2 *Plot the posterior samples for α and β. What is the relationship between the two variables, and why?*

Problem 2.3 *By using the "generated quantities" code block or otherwise generate samples from the posterior predictive distribution. By overlaying the real time series for each individual on a graph of the posterior predictive comment on the fit of the model to data.*

Problem 2.4 *Fit a model with separate (α, β) for each individual in the dataset. Use separate and independent normal(0,250) priors for the parameters. Again use 1000 samples per chain over 4 chains.*

Problem 2.5 *Compute the posterior mean estimates of the β parameters for the new "heterogeneous-parameters" model. How do these compare to the single β estimate obtained for the homogeneous model?*

Problem 2.6 *Using the "generated quantities" code block, or otherwise, generate samples from the posterior predictive distribution. By comparing individual subject data to the posterior predictive samples, comment on the fit of the new model.*

Problem 2.7 *Partition the data into two subsets: a training set (of subjects 1-17) and a testing set (of subject 18 only). By fitting both models - the heterogeneous and homogeneous coefficients models - on the training sets, compare the performance of each model on predicting the test set data.*

Problem 2.8 *Alternatively we can fit a hierarchical model to the data which (hopefully) captures some of the best elements of each of the aforementioned models. Here we assume that the individual* (α, β) *for each subject are allowed to vary, but there is some overarching "population-level" distribution from which they are drawn. Assume that the coefficients have the following relationships:*

$$\alpha_i \sim normal(a, b) \tag{12}$$
$$\beta_i \sim normal(c, d) \tag{13}$$
$$a \sim normal(100, 100) \tag{14}$$
$$b \sim cauchy(0, 5) \tag{15}$$
$$c \sim normal(10, 5) \tag{16}$$
$$d \sim cauchy(0, 1) \tag{17}$$

Code up the above model and compare the posterior distribution for β *for the hierarchical model, with those from the heterogeneous ones.*

Problem 2.9 *Graph the posterior distribution for* β *for another individual (not in the original dataset). How does this distribution compare to the value of* β *obtained from the homogeneous coefficient model? (Hint: use the "generated quantities" block to generate samples of* β *from the top-level parameters c and d.)*

3 Hierarchical ODEs: Bacterial cell population growth

The file "prob6_odeData.csv" contains data for 5 replicates of an experiment in which bacterial cell population numbers were measured over time (as in the lectures). The following model for bacterial population size is proposed to explain the data:

$$\frac{\mathrm{d}N}{\mathrm{d}t} = \alpha N(1 - \beta N) \tag{18}$$

However measurement of bacterial cell numbers is subject to random, uncorrelated measurement error:

$$N^*(t) \sim \text{normal}(N(t), \sigma) \tag{19}$$

where $N^*(t)$ is the measured number of cells, and $N(t)$ is the true population size. Finally we suppose that the initial number of bacteria cells is unknown, and hence must be estimated.

Further we assume the following priors:

$$\alpha \sim normal(0, 2)$$
$$\beta \sim normal(0, 2)$$
$$\sigma \sim cauchy(0, 1)$$
$$N(0) \sim normal(5, 2)$$

where all parameters have a lower value of zero.

Problem 3.1 *Write a Stan function that returns* $\frac{\mathrm{d}N}{\mathrm{d}t}$*. (Hint 1: this will need to be done within the "functions" block at the top of the Stan file. Hint 2: the function must have a structure:*

```
real[] bacteria_deriv(real t,real[] y,real[] theta,real[] x_r,int[] x_i)
```

where the variables x_i *and* x_r *are not used here, but nonetheless need to be defined:*

```
transformed data {
    real x_r[0];
    int x_i[0];
}
)
```

Problem 3.2 *Estimate a model where the parameters* (α, β) *are assumed to be the same across all experimental replicates.*

Problem 3.3 *By graphing the data, or otherwise, comment on the assumption of a common* (α, β) *across all replicates.*

Problem 3.4 *Now estimate a model that estimates separate values for* (α, β) *across all replicates. Graph the posterior distribution for each parameter.*

Problem 3.5 *Estimate a hierarchical model assuming the following priors:*

$$\alpha \sim gamma(a, b)$$
$$\beta \sim gamma(c, d)$$
$$a \sim normal(20, 5)$$
$$b \sim normal(40, 5)$$
$$c \sim normal(10, 3)$$
$$d \sim normal(100, 5)$$

Compare your estimates of (α, β) *with those from the completely heterogeneous model.*

Problem 3.6 *Estimate the overall* (α, β) *for the hierarchical model. How do these compare to the pooled model estimates?*

Problem 3.7 *By holding out one of your datasets, compare the predictive performance of each model.*

4 Bowel cancer model selection

The file "prob6_cancer.csv" contains (fictitious) data on the population size of a given county (N) and the number of bowel cancer cases in that county (X). In this question we aim to build a model to estimate the underlying rate of cancer occurrence λ.

Problem 4.1 *A simple model is to assume that cancer occurrence is an independent event, and hence we use the following model,*

$$X_i \sim poisson(N_i\lambda) \tag{20}$$

where N_i is the population in county i, and X_i is the number of cases of bowel cancer in the same county. In Stan write a model to estimate the underlying rate of bowel cancer occurrence (λ), where we assume a prior of the form $\lambda \sim normal(0.5, 0.5)$.

Problem 4.2 *Using the* `generated quantities` *section record the estimated log likelihood of each data point, for each posterior sample of λ.*

Problem 4.3 *By using Stan's* `optimizing` *function to obtain the MAP estimate of λ, estimate the expected log pointwise predictive density (elpd) via a DIC method,*

$$\widehat{elpd} = log\ p(X|\hat{\theta}_{Bayes}) - \underbrace{2V_{s=1}^{S} log\ p(X|\theta_s)}_{DIC} \tag{21}$$

where $V_{s=1}^{S} log\ p(X|\theta_s)$ is the variance in log-likelihood for all data points across S posterior draws. Hint: the latter part of the formula requires that we estimate the model by sampling.

Problem 4.4 *Estimate elpd using the AIC method. Hint: use Stan's* `optimizing` *function where the Stan file has had the prior commented out, to achieve the maximum likelihood estimate of the log-likelihood.*

Problem 4.5 *Either manually or using the "loo" package in R estimate the elpd by a WAIC method. If you choose the manual method, this can be done with the following formula,*

$$\widehat{elpd} = \underbrace{\sum_{i=1}^{N} log\left(\frac{1}{S}\sum_{s=1}^{S} p(X_i|\theta_s)\right)}_{log\ pointwise\ predictive\ density} - p_{WAIC} \tag{22}$$

where $p_{WAIC} = \sum_{i=1}^{N} V_{s=1}^{S} var_{post}\left[log\ p(X_i|\theta_s)\right]$.

Problem 4.6 *By partitioning the data into 10 folds of training and testing sets (where one data point occurs in each testing set once only), estimate the out-of-sample predictive capability of the model. Hint 1: in R use the "Caret" package's* `createFolds` *to create 10 non-overlapping folds. Hint 2: adjust your Stan program to calculate the log-likelihood on the test set.*

Problem 4.7 *A colleague suggests fitting a negative binomial sampling model to the data, in case over-dispersion exists. Using a prior* $\kappa \sim$ *log-normal*$(0, 0.5)$ *on the dispersion parameter, change your Stan model to use this distribution, and estimate the out-of-sample predictive density using any of the previous methods. Which model do you prefer? Hint: use Stan's* `neg_binomial_2` *function to increment the log-probability.*

Problem 4.8 *A straightforward way to estimate the marginal likelihood is to use,*

$$p(X) \approx \frac{1}{S} \sum_{s=1}^{S} p(X|\theta_s) \tag{23}$$

where $\theta_s \sim p(\theta)$*. Either using Stan's* `generated quantities` *block or otherwise estimate the marginal likelihood of the poisson model. Hint: if you use Stan then you need to use* `log_sum_exp` *to marginalise the sampled log probabilities.*

Problem 4.9 *Estimate the marginal likelihood of the negative binomial model, and hence estimate the log Bayes Factor. Which model do you prefer?*

References

[1] Gregory Belenky, Nancy J Wesensten, David R Thorne, Maria L Thomas, Helen C Sing, Daniel P Redmond, Michael B Russo, and Thomas J Balkin. Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: A sleep dose-response study. *Journal of sleep research*, 12(1):1–12, 2003.

[2] John B Carlin. Meta-analysis for 2× 2 tables: A bayesian approach. *Statistics in medicine*, 11(2):141–158, 1992.