Course project for "Bayesian Data Analysis"

Formalities

The course project consists of four parts. You need to answer part I and III and *either part II or IV*. Note that questions 6 and 7 in part III are not mandatory. You are welcome to answer all four parts if you wish :-) Hand in your answers no later than June 26. Include relevant figures, and please also include the important parts of your R code. If you hand in electronically, please collect all parts of the report to *one pdf file*. In other words: no word files, only one document.

Data

The data consists of the number of failures for 10 pump stations at a nuclear power plant. Pump station i was under observation in a time interval of length t_i (thousands of hours), and y_i failures were observed during that time interval. The data is listed in the table.

Pump i	Failures y_i	Time t_i
1	5	94.320
2	1	15.720
3	5	62.880
4	14	125.760
5	3	5.240
6	19	31.440
7	1	1.048
8	1	1.048
9	4	2.096
10	22	10.480

In parts I and II you are asked to make Bayesian analyses of the observed failure rates, $r_i = y_i/t_i$, whereas in parts III and IV you are asked to use a hierarchical model of the observed number of failures (the y's).

Part I: Analysis of observed failure rates with normal models

Consider first a normal model for $r_1, \ldots, r_n$, with a non-informative prior on the parameters.

1. Specify the model in details and make 1000 simulations from the posterior distribution of the parameterers.
2. Make simulations from the predictive posterior distribution (the posterior distribution of future observations) and report the mean. Explain why the model is not a good one.

Consider instead the logarithmic failure rates, $z_i = \log(r_i)$ as your response, and assume that the z_i's are iid. $N(\mu, \sigma^2)$. Put a non-informative prior on (μ, σ^2). Note that you can probably re-use much of your R-code from the previous questions.

4. Make simulations from the posterior distribution of the parameters.

5. Compute the median, the 5% and the 95% quantile for the posterior distribution of e^{μ}. What is the interpretation of e^{μ}? And why, do you think, do I ask for the median rather than the mean?

6. Make simulations from the predictive posterior distribution of r and compare to the data. Does this model seem to more appropriate?

Part II: Analysis of observed failure rates with an exponential model

As an alternative, consider now an exponential model for the failure rates: Assume that, conditional on γ, $r_1, \ldots, r_n$ are independent and exponential with rate parameter γ. Hence,

$$p(r|\gamma) = \prod_{i=1}^{n} \gamma e^{-\gamma r_i}$$

Recall that the Gamma-distribution with shape parameter $\alpha > 0$ and inverse scale parameter $\gamma > 0$ has density

$$f(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad x > 0$$

1. Assume a flat prior for γ on $(0, \infty)$. What is the corresponding posterior (and is it at all proper?). Make 1000 simulations from the posterior distribution of γ.

2. What is the median in the conditional distribution of r given γ (as a function of γ). Compute the median, the 5% and the 95% quantile in the posterior distribution of this median. Compare to question 5 from part I.

3. Show that the Gamma distribution is conjugate for the exponential model. In particular, what is the posterior distribution of γ if the prior is a Gamma distribution with shape α and inverse scale β? What is the posterior mean? For fixed prior parameters (α, β), what happens to the posterior mean as the number of observations (n) increases?

Part III: A hierarchical model

We will now make an analysis of the data based on a hierarchical model. In the following, let $\lambda = (\lambda_1, \ldots, \lambda_n)$ denote the collection of 'true' (but unobserved) failure rates for the pump stations.

The model has three levels:

- Conditional on the failure rates λ, the number of failures (y_i) are independent and y_i is Poisson with mean $\lambda_i t_i$.
- Conditional on the hyperparameter β, the pump failure rates (λ_i) are assumed to be iid. exponential with rate β.
- The hyperparameter β is exponential with rate 1.

And now the questions:

1. Explain why this might be a reasonable model. What are the unknowns in the model? Work out the joint posterior of all the unknowns.

2. Work out the conditional posterior of β given λ as well as the conditional posterior of λ given β. Describe how the Gibbs sampler would work in this situation.

3. Implement the Gibbs sampler in R (without the use of BUGS) and investigate its convergence properties. Note that you only need initial for values for λ (or β); then the Gibbs sampler takes care of β (or λ).

4. Report the means (or medians) of the posterior distributions of the individual pump failure rates and compare to the naive observed failure rates y_i/t_i.

5. The rate 1 in the hyperprior distribution was chosen ad hoc. Investigate how sensitive the posterior results are to this prior rate. In particular, for which pump stations are the failure rates most sensitive? Try to explain your observations.

6.* [This question is not mandatory] Assume that we repeat the experiment with the *same pumps* and the *same time period lengths*. Make simulations from the posterior predictive distribution and report the predictive posterior mean of the number of failures for each pump.

7.* [This question is not mandatory] Finally assume that we repeat the experiment with 10 *new pumps* but with the *same time period lengths* as in the original experiment. Make simulations from the posterior predictive distribution and report the predictive posterior mean of the number of failures for each pump. Comment on the model's ability to generate data like the observed ones.

Part IV: BUGS

1. Implement the hierarchical model i BUGS and make sure that you get the same results as you did with your own Gibbs sampler from part III.