Bayesian Analysis: Practical Sessions

Session 4: Introduction to JAGS

Goals:

- Implement a model with JAGS

Exercise 4.1 Suicides: Estimating a proportion. A Danish hospital wants to investigate the reincidence of attempted suicide by poisoning. Although there are studies about that made in USA, Danish researchers consider that their conclusions do not apply to Denmark. In order to estimate the percentage of reincidence, they monitored 974 patients that attempted suicide during the 10 years that followed their first attempt, and found out that 103 attempted suicide a second time. They also lost 100 of the initial patients to follow up, but the researchers are confident that the missing data are random, meaning that the probability of a second suicide attempt is not related to loss to follow up.

a) Give a point estimate of the probability to commit suicide again in the next ten years.
b) Predict the number of suicides on patients who were lost to follow up, and give a credibility interval of the prediction of 95% of probability.

Exercise 4.2 Systolic blood pressure: Estimating the mean of a normal location model. It has been established that the standard deviation of the systolic blood pressure everywhere in the world is 13, but its mean varies slightly from country to country around an overall worldwide mean equal to 125. The department of Health of Andorra wants to:

a) compute a point estimate for the mean of the systolic blood pressure of the inhabitants of Andorra, μ,
b) predict the systolic blood pressure of one resident of Andorra,
c) determine whether the mean of the systolic blood pressure of Andorrans is higher than the overall world mean of the systolic blood pressure.

For this purpose, they obtain a sample of the systolic blood pressure of twenty Andorrans, which is: 98, 160 , 136 , 128 , 130 , 114 , 123 , 134 , 128 , 107, 123 , 125 , 129 , 132 , 154 , 115 , 126 , 132 , 136 and 130 .

Exercise 4.3 Lightning in a storm: estimating a frequency rate. The meteorological observatory of "Turó de l'home" has collected, among other information, the number of lightning in a given storm together with the duration of that storm in minutes. The data corresponding to a stormy summer is presented in the following table:

Lightning	2	10	0	0	12	2	6	8	11
Duration	20	30	25	10	60	25	40	35	25

Lightning	0	27	2	0	10	2	1	3	1
Duration	40	45	25	10	10	15	25	50	20

a) Estimate the lightning frequency using a conjugate prior distribution.
b) Estimate the lightning frequency using a non-conjugate prior distribution.
c) Using the prior in b), estimate the probability that the number of lightning during half an hour of a storm is 0.

Exercise 4.4 Two methods of training workers: Comparing two means. The Human Resources Department of a large company wishes to compare two methods of training industrial workers to perform a skilled task. Twenty workers are selected: 10 of them are randomly assigned to be trained using method A, and the other 10 are assigned to be trained using method B. After the training is complete, all the workers are tested on the speed of performance at the task. The times taken to complete de task are:

Method A	Method B
115	123
120	131
111	113
123	119
116	123
121	113
118	128
116	126
127	125
129	128

a) We assume that the observations come from *Normal*(μ_A,σ) and *Normal*(μ_B,σ), where σ=6. Use independent *Normal(m,s)* prior distributions for μ_A and μ_B respectively, where m=100 and s=20. Find the posterior distributions of μ_A and μ_B.
b) Find the posterior distribution of $\mu_A - \mu_B$.
c) Find a 95% Bayesian credible interval for $\mu_A - \mu_B$.
d) Repeat parts a), b) and c) supposing that σ is unknown.

Exercise 4.5 Twin cows: Comparing two means between paired observations. An experiment was designed to determine whether a mineral supplement was effective in increasing annual yield in milk. Fifteen pairs of identical twin dairy cows were used as the experimental units. One cow from each pair was randomly assigned to the treatment group that received the supplement. The other cow from the pair was assigned to the control group that did not receive the supplement. The annual yields are recorded below:

Twin Set	Milk Yield: Control (liters)	Milk Yield: Treatment (liters)
1	3525	3340
2	4321	4279
3	4763	4910
4	4899	4866
5	3234	3125
6	3469	3680
7	3439	3965
8	3658	3849
9	3385	3297
10	3226	3124
11	3671	3218
12	3501	3246
13	3842	4245
14	3998	4186
15	4004	3711

Assume that the annual yields from cows receiving the treatment are $Normal(\mu_t,\sigma_t)$, and the annual yields from the cows in the control group are $Normal(\mu_c,\sigma_c)$. The cows in the same pair share identical genetic background, their responses will be more similar than two cows that were from different pairs. There is natural pairing. As the samples drawn from the two populations cannot be considered independent of each other, they decided to take differences $d_i = y_{i1} - y_{i2}$. The differences bill be $Normal(\mu_d,\sigma_d)$, where $\mu_d = \mu_t - \mu_c$. To determine whether or not the treatment was effective in increasing the yield of milk, you would perform the one-sided hypothesis test:

$$H_1 : \mu_d \leq 0 \quad \text{vs} \quad H_2 : \mu_d > 0 .$$

Exercise 4.6 Linear Regression. Brain weight. Build a linear model to try to explain the brain weight of a mamal as a function of its body weight through the data for 62 mammals and give a credible interval for the brain weight of an animal whose body weight is 100 Kg.

The data is in *Brain.txt* (the body weight is measured in kilograms and the brain weight in grams).

Exercise 4.7 Yield of Potatoes. A researcher is investigating the relationship between yield of potatoes (y) and level of fertilizer (x). She divides a field into eight plots of equal size and applied fertilizer at a different level to each plot. The level of fertilizer and yield for each plot is recorded below:

Fertilizer level x	Yield y
1	25
1.5	31
2	27
2.5	28
3	36
3.5	35
4	Not available
4.5	34

Suppose that we know that yield given the fertilizer level is $Normal(\beta_0 + \beta_1 x, \sigma)$.

a) Using non-informative priors for the parameters find the posterior distribution of β_1.
b) Find a 95% credible interval for y given x=4.

Exercise 4.8 Final grades from Bayesian analysis course. The lecturers of the Bayesian analysis course want to know if the grades get by the boys and the girls are equal. The final grades are in the next table:

boys	girls
9.6	6.1
7.0	9.1
5.0	8.8
8.0	5.7
8.4	8.9
6.4	6.1
	6.5

The lectures have chosen this statistical model:

$$y|\ \beta_0, \beta_1, \sigma \sim Normal(\beta_0 + \beta_1 sex, \sigma),$$

where *sex* is a dichotomy variable which equals 1 for the boys and 0 for the girls.

Answer the following questions:

a) Write the bayesian model and justify the chosen prior distribution.
b) Update the model and draw the posterior distribution for every parameter.
c) Do you think there are differences between the grades get by the boys and the grades get by the girls?
d) Modify the model in order to capture or evaluate the fact that the within variability in boys could be different in girls.

Exercise 4.9 Multiple linear Regression. In the file *graduates.txt* you will find a sample of 22 universities. For each one there are the percentage of graduate students in 6 years, the average of SAT, the average price per student and the type of school (1 boys and girls are segregated, 0 boys and girls are mixed). Build a Bayesian model to explain the relationship between the percentage of graduate students and the other covariables; and interpret it.

Exercise 4.10 Leukemia: Time to Event Data. Feigl and Zelen (1965) present data on the survival times in weeks of patients who were diagnosed with leukemia. The patients were classified according to one characteristic of white cells referred to as AG+ and AG-. The n_1=17 times from diagnosis to death for the AG+ group are: 65, 156, 100, 134, 16, 108, 121, 4, 39, 143, 56, 26, 22, 1, 1, 5, 65, and the n_2=16 observations for the AG- group are: 56, 65, 17, 7, 16, 22, 3, 4, 2, 3, 8, 4, 3, 30, 4, 43. No prior information is available. Suppose a two-sample exponential model for handling the leukemia data.

a) Draw the survival function for every treatment jointly with their 95% credible interval.
b) Calculate a 95% credible interval for the difference of the 24-week (approximately 6-month) probabilities of survival for the two groups.

Exercise 4.11 Weight and height. In the file WeightHeight.txt there are data about the weight, height and sex of several students. Implement the next models:

a) A model where the weight is explained as a function of the height.
b) A model where the weight is explained as a function of the height and the sex.
c) A model where the weight is explained as a function of the height, the sex and their interaction.

Exercise 4.12 basketball. The objective is to carry out an analysis to compare the number of points that are scored in a NBA's match with the number of points that are scored in an ACB's match. In the data file *Basquet.txt* you will find the total number of points of 20 matches taken randomly from ACB and 20 taken from NBA. Estimate the probability that the points scored in a NBA match will be 30 points larger than the points scored in a ACB match:

a) Assuming a Poisson model for the number of points in a match.

b) Assuming a Normal model (as an approximation) for the number of points in a match.

c) Think about the differences in using different statistical models.