Course EPIB-675 - Bayesian Analysis in Medicine

Assignment 8

1. Suppose you have observed 5 successful surgical outcomes in 7 trials.

(a) What is your posterior distribution and 95% interval if you start with a Jeffrey's prior (beta(0.5, 0.5))? What is your posterior distribution and 95% interval if you start with a Haldane prior (beta(0, 0))? What is your posterior distribution and 95% interval if you start with a uniform prior (beta(1, 1))?

(b) Use R to graph the three posterior distributions you found in part (a). Put them all on the same graph, and label which is which using the "legend" command.

(c) Comment on the degree to which the posterior distributions differ from each other. Is the particular choice of "non-informative" prior important?

2. Repeat parts (a), (b), and (c) of question 1, but now suppose that you have a bigger data set, with 50 successes in 70 trials.

3. Consider two researchers who start with different opinions about the success rate of a new treatment. Researcher A believes that the success rate will be near 50%, and decides that a beta(50, 50) prior distribution accurately represents his prior views. Researcher B is much more optimistic, and states that his prior is a beta(80, 20).

(a) Graph these two prior distributions in R.

(b) Suppose that the true success rate is in fact 80%. Suppose that a clinical trial will be carried out, and the trial planners want to know how much evidence will be needed so that the two investigators will end up with very similar posterior distributions, as measured by the 95% intervals. What sample size would be required to ensure that their upper and lower intervals do not differ by more than 1% each? You can assume that the data collected will have observed success rate exactly equal to the true rate. In other words, if sample size is 100, there will be 80 successes, if sample size is 200 there will

be exactly 160 successes, etc. [Hint: Unless you have very good programming skills, you will likely need to do this by trial and error.]

4. In the article by Raftery on model selection that is in the course pack, he claims that backwards and forwards regression leads to models that are generally too large. In this question, we will empirically investigate if this tends to be true, and also see if Bayes Factors as approximated by the BIC leads to a better solution.

(a) Generate a 100 by 25 matrix of simulated normal(0,1) random noise data and save it as a data frame, by using the following R command:

```
q4.data <- data.frame(matrix(rnorm(2500), nrow=100, byrow=T))
```

We will assume that the first column of the matrix is our dependent variable, and the rest of the columns are potential predictor variables. Of course, since all we have is random noise, the "best" model should just be the null model, i.e., with just an intercept (and the intercept should be estimated to be near zero, the overall expected mean of the data in column 1). Use the R command

```
summary(lm(q4.data[,1] ~ q4.data[,2] + q4.data[,3] + q4.data[,4] +
q4.data[,5] + q4.data[,6] + q4.data[,7] + q4.data[,8] +
q4.data[,9] + q4.data[,10] + q4.data[,11] + q4.data[,12] +
q4.data[,13] + q4.data[,14] + q4.data[,15] + q4.data[,16] +
q4.data[,17] + q4.data[,18] + q4.data[,19] + q4.data[,20] +
q4.data[,21] + q4.data[,22] + q4.data[,23] + q4.data[,24] +
q4.data[,25]) )
```

to look at regression of the first column of the matrix of random numbers on the next 24 columns. How many of them have p-values below 0.05? [Note: Since we are generating random data, each student's answers should be slightly different, but there should be overall trends apparent].

(b) Now use backwards model selection with criterion AIC to find a "final model". You can use an R command such as

```
step(lm(q4.data[,1] ~ q4.data[,2] + q4.data[,3] + q4.data[,4] +
```

```
q4.data[,5] + q4.data[,6] + q4.data[,7] + q4.data[,8] +
q4.data[,9] + q4.data[,10] + q4.data[,11] + q4.data[,12] +
q4.data[,13] + q4.data[,14] + q4.data[,15] + q4.data[,16] +
q4.data[,17] + q4.data[,18] + q4.data[,19] + q4.data[,20] +
q4.data[,21] + q4.data[,22] + q4.data[,23] + q4.data[,24] +
q4.data[,25]), direction="backward")
```

Note: Output will be very long, but we are just interested in the very last couple of lines, that give the final model according to the AIC.

(c) We will now compare the answer to a Bayesian approach using the BIC criterion. First, you need to download the BMA (Bayesian Model Averaging) program, which is found here:

`http://cran.r-project.org/src/contrib/Descriptions/BMA.html`

The BMA package contains many functions, we will just use bicreg, which performs Bayesian model selection for linear regression, to find a final model (i.e., the model that is "best" according to the BIC criterion). There are instructions for how to use this function in the first few lines of the function itself, or look at the html help in R. How does it compare to the models suggested in part (b)?

5. In this question we again will use the bicreg program to examine various models relating to predicting IQ.

(a) Obtain the data file "brain.data.txt" from the course web page. Variables used are:

CCMIDSA:	Corpus Collasum Surface Area (cm2)
IQ:	Intelligence Quotient
HC:	Head Circumference (cm)
ORDER:	Birth Order
SEX:	Sex (1=Male 0=Female)
TOTSA:	Total Surface Area (cm2)
TOTVOL:	Total Brain Volume (cm3)
WEIGHT:	Body Weight (kg)

Read the data set into R (for example, using a "read.table" command, note that a header is present, so use "header=T"), and use the "pairs" command

to produce a pairwise plot. Comment on which variables seem to be related or not, just looking at the graphs alone.

(b) Use the bicreg program to perform model selection on this data set. Provide the output in your answer.

(c) Look at the best model. How much better is it than the second best model?