

ST440/540 Applied Bayesian Analysis

Lab activity for 3/25/2024

A. HOMEWORK AND QUIZ SOLUTIONS

Q7: In Bayesian multiple linear regression with known error variance the Jeffries prior for the regression coefficients is prior $\pi(\beta) \propto 1$. Does this always lead to a proper posterior? Why or why not?

The posterior is not always proper. It is proper if $X'X$ is invertable, which requires $n > p$ and the covariates to be linearly independent.

Chapter 4, Problem 2

(2a)

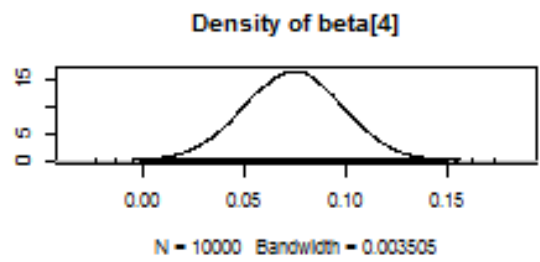
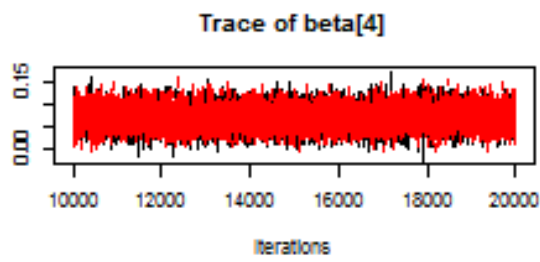
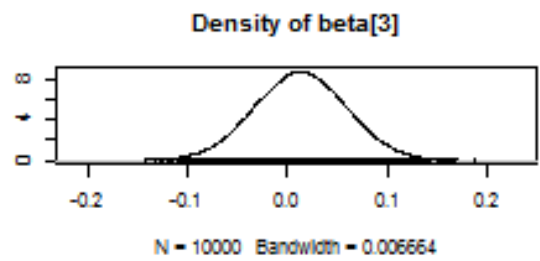
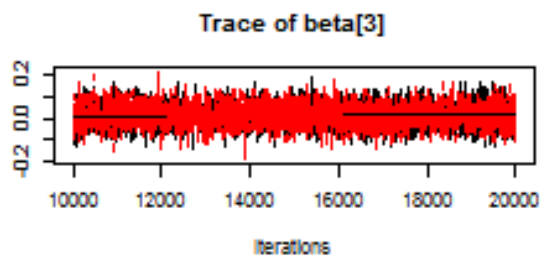
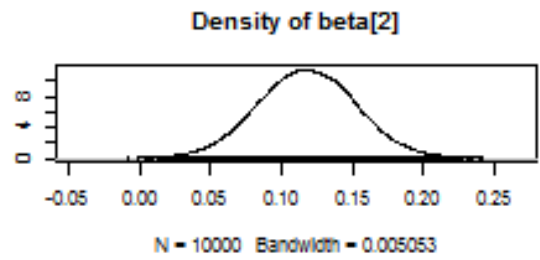
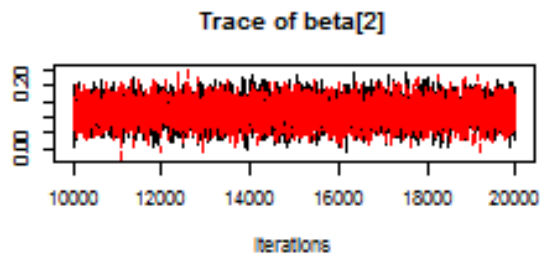
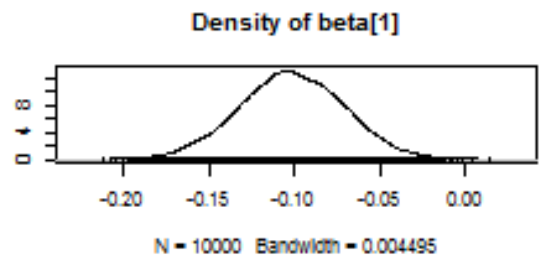
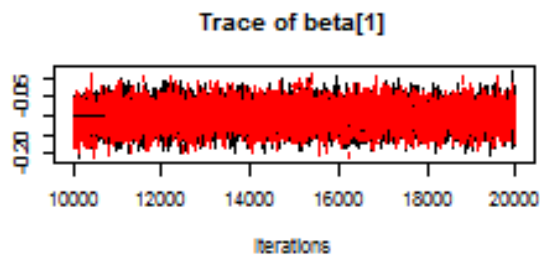
```
library(MASS)
data(Boston)
X <- scale(Boston[,1:13])
Y <- as.vector(scale(Boston[,14]))

library(rjags)
data <- list(n=length(Y), p=ncol(X), X=X, Y=Y)

model_string <- textConnection("model{

  # Likelihood
  for(i in 1:n){
    Y[i] ~ dnorm(mu[i], tau)
    mu[i] <- alpha + inprod(X[i,], beta[])
  }
  for(j in 1:p){
    beta[j] ~ dnorm(0, 0.01)
  }
  alpha ~ dnorm(0, 0.01)
  tau ~ dgamma(0.1, 0.1)
}")

model <- jags.model(model_string, data = data, n.chains=2, quiet=TRUE)
update(model, 10000, progress.bar="none")
params <- c("beta")
samples <- coda.samples(model,
  variable.names=params,
  n.iter=10000, progress.bar="none")
plot(samples)
```



```
sum1 <- summary(samples)$stat[,1:2]  
rownames(sum1) <- colnames(X)
```

```

round(sum1,3)
##           Mean      SD
## crim    -0.101 0.031
## zn       0.118 0.035
## indus    0.013 0.046
## chas     0.074 0.024
## nox     -0.223 0.048
## rm       0.292 0.032
## age      0.002 0.041
## dis     -0.338 0.046
## rad      0.287 0.063
## tax     -0.222 0.069
## ptratio -0.224 0.031
## black    0.093 0.027
## lstat   -0.407 0.040

```

Convergence looks great. All covariates except age and indus have 95% intervals that exclude zero.

(2b)

```

sum2 <- summary(lm(Y~X))$coef[,1:2]
round(sum2,3)
##           Estimate Std. Error
## (Intercept)    0.000      0.023
## Xcrim          -0.101      0.031
## Xzn             0.118      0.035
## Xindus          0.015      0.046
## Xchas           0.074      0.024
## Xnox           -0.224      0.048
## Xrm             0.291      0.032
## Xage            0.002      0.040
## Xdis           -0.338      0.046
## Xrad            0.290      0.063
## Xtax           -0.226      0.069
## Xptratio       -0.224      0.031
## Xblack          0.092      0.027
## Xlstat         -0.407      0.039

```

The results are nearly identical to the Bayesian analysis with uninformative priors, as expected.

(2c)

```

model_string <- textConnection("model{

  # Likelihood
  for(i in 1:n){
    Y[i] ~ dnorm(mu[i],tau)
    mu[i] <- alpha + inprod(X[i,],beta[])
  }
  for(j in 1:p){
    beta[j] ~ ddexp(0,taub)
  }
  alpha ~ dnorm(0, 0.01)
  tau ~ dgamma(0.1, 0.1)
  taub ~ dgamma(0.1, 0.1)
}")

model <- jags.model(model_string,data = data, n.chains=2,quiet=TRUE)
update(model, 10000, progress.bar="none")
params <- c("beta")
samples <- coda.samples(model,
  variable.names=params,
  n.iter=10000, progress.bar="none")
sum3 <- summary(samples)$stat[,1:2]
rownames(sum3) <- colnames(X)
round(sum3,3)
##           Mean      SD
## crim    -0.093 0.031
## zn       0.106 0.035
## indus   -0.001 0.042
## chas     0.074 0.024
## nox     -0.204 0.047
## rm       0.295 0.032
## age     -0.002 0.037
## dis     -0.322 0.045
## rad      0.244 0.065
## tax     -0.184 0.069
## ptratio -0.218 0.031
## black    0.090 0.027
## lstat   -0.406 0.039

```

In this case with $n \gg p$ the results of the Bayesian lasso are similar to those from the analysis with uninformative priors.

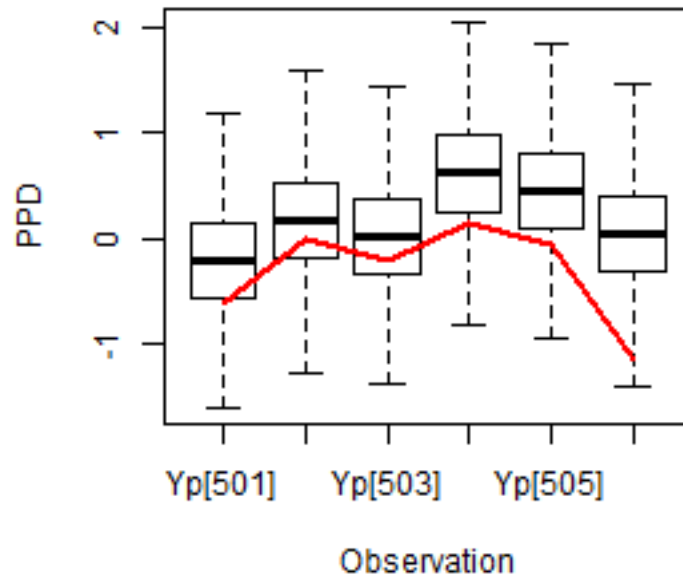
(2d)

```

model_string <- textConnection("model{
  # Likelihood
  for(i in 1:500){
    Y[i] ~ dnorm(mu[i],tau)
    mu[i] <- alpha + inprod(X[i,],beta[])
  }
  for(j in 1:p){
    beta[j] ~ dnorm(0,0.01)
  }
  alpha ~ dnorm(0, 0.01)
  tau ~ dgamma(0.1, 0.1)
  for(i in 501:n){
    Yp[i] ~ dnorm(mup[i],tau)
    mup[i] <- alpha + inprod(X[i,],beta[])
  }
}")

model <- jags.model(model_string,data = data, n.chains=2,quiet=TRUE)
update(model, 10000, progress.bar="none")
params <- c("Yp")
samples <- coda.samples(model,
  variable.names=params,
  n.iter=10000, progress.bar="none")
Yp <- rbind(samples[[1]],samples[[2]])
boxplot(Yp,outline=FALSE,xlab="Observation",ylab="PPD")
lines(Y[501:506],lwd=2,col=2)

```



The observed values all fall in center of the PPD.

B. DISCUSSION QUESTIONS

(1) A study randomized 100 subjects to either treatment or control groups. The treatment group will take a high-intensity spinning class each morning and the control group will continue their normal routine. Each patient will have their blood pressure measured at baseline and once a week for each of the four weeks of the study. The goal is to determine whether spinning reduces blood pressure.

(a) Describe a model and priors for these data

As with all of these questions, there is no single right answer here. In reality, you would first explore the data a bit, then try and compare a few models before settling on a final analysis. Below are just some ideas to get started. Let Y_{ij} be the BP for subject i at time $j = 0, 1, 2, 3, 4$. We could allow each subject to have a separate linear regression, so

$$Y_{ij} = a_i + b_i * j + e_{ij}$$

The random intercept for subject i has distribution $a_i \sim N(m, s_a^2)$. The slope depends on treatment group $x_i = 1$ if spin, $x_i=0$ if not, so $b_i \sim N(c+d*x_i, s_b^2)$. To complete the Bayesian model we could pick conjugate priors $m, c, d \sim N(0, 100)$ and $s_a^2, s_b^2 \sim \text{InvGamma}(0.1, 0.1)$.

(b) How would you summarize the results?

The assumption made here is that taking spin class doesn't affect the BP at baseline (a_i), but affects the slope of BP (b_i) once the study starts. So, if $d < 0$ then spin class lowers the average slope, i.e., generally improves BP. I would compute the 95% interval for d and if it excludes zero conclude spin affects BP.

(c) Write JAGS code for this model

```
for(i in 1:100){for(j in 1:5){
  Y[i,j] ~ dnorm (a[i] + b[i]*j, sigy2inv)
}}
for(i in 1:100){
  a[i] ~ dnorm(m, siga2inv)
  b[i] ~ dnorm(c + d*X[i], sigb2inv)
}
#priors...
```

(2) A group of 10 ecologists is surveying a forest for red cockaded woodpeckers (RCP). Each ecologist will walk along a different path and make 5 stops. At each stop, they will record local conditions (tree density, elevation, etc.) and whether they see or hear an RCP. The objective is to build a model for the types of habitat that are the most favorable to the RCP.

(a) Describe a model and prior for these data

$Y_{ij} = 1$ if ecologist i hears an RCP on stop j and $Y_{ij}=0$ otherwise. The response is binary and clustered by ecologist, so a mixed effects logistic regression model would be appropriate.

$$\text{Logit}[\text{Prob}(Y_{ij}=1)] = b_0 + b_1 * \text{tree_density}_{ij} + b_2 * \text{elevation}_{ij} + a_i$$

where the random effects have distribution $a_i \sim \text{Normal}(0, s^2)$. Uninformative priors are $b_0, b_1, b_2 \sim \text{Normal}(0, 10^2)$ and $s^2 \sim \text{InvGamma}(0.1, 0.1)$.

(b) How would you summarize the results?

We can test for covariate effects by comparing the posteriors of b_1 and b_2 to zero.

(c) Write JAGS code for this model

```
for(i in 1:10){for(j in 1:5){
  Y[i,j] ~ dbern(p[i,j])
  logit(p[i,j]) = b0 + b1*tree_density[i,j] + b2*elevation[i,j] + a[i]
}}
for(i in 1:10){
  a[i] ~ dnorm(0, sigma2inv)
}
#priors...
```

(3) In a study of the genetic determinants of smoking addiction, researchers sampled 1,000 people and asked whether they smoked. For each subject, they also recorded 10,000 genetic markers. The objective is to determine if any of the markers are associated with smoking addiction.

(a) Describe a model and prior for these data

Since the response is binary and there are many covariates, we could fit a logistic regression with Bayesian LASSO prior for the regression coefficients.

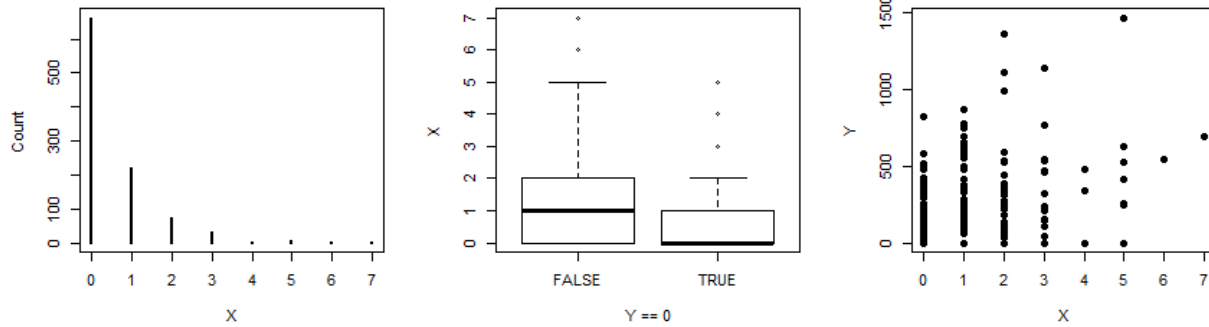
(b) How would you summarize the results?

I would conclude markers are important if their 95% posterior credible set excluded zero.

(c) Write JAGS code for this model

See listing 4.4 in the book.

(4) To study the effect of an online advertising campaign by your company, you gather data for 1,000 consumers and record the number of ads they have been exposed to (X) and the amount of money (\$) they spent on your product in the past year (Y). The data are plotted below.



(a) Describe a model for these data

The response variable Y is a mix of zeros and positive continuous response. So, I would build two models, one a logistic regression for whether the response is or is not zero, and then a log-linear regression for the non-zeros. So

$$\text{logit}(\text{Prob}(Y>0)) = a_0 + a_1 X \text{ and for } Y>0, \log(Y) \sim \text{Normal}(b_0 + b_1 X, s^2).$$

For priors, $a_0, a_1, b_0, b_1 \sim \text{Normal}(0, 10^2)$ and $s^2 \sim \text{InvGamma}(0.1, 0.1)$.

(b) How would you summarize the results?

The covariate affects the response if either a_1 or b_1 is non-zero.

(c) Write JAGS code for this model

Since the model for zero/non-zero and $\log(Y)$ do not share any parameters, you could run JAGS twice, separately. First you would use all observations and let the response be $Z_i = 1$ if $Y_i > 0$ and $Z_i = 0$ if $Y_i = 0$. This would be standard Bayesian logistic regression. Second you could run JAGS using only the observations with $Y_i > 0$ and take the response as $\log(Y_i)$. This would be standard Bayesian multiple regression.

(5) Consider the class participation question about sports and missed days.

(a) From the analysis in my answer, why can't we conclude that playing sports causes an improvement in attendance?

There could be a missing confounder (lurking variable). A variable is a confounder if it is correlated with both X and Y. So maybe playing sports has nothing to do with missing school, but SES is related to both attendance and sports and is thus a confounder. If you don't include SES in the model, you would find an association between sports and attendance. But this would not be a causal relationship, but just an artifact of SES' effect on both variables.

(b) How would we have to modify the analysis to have a hope of establishing a causal relationship?

You could add SES as a covariate, so

$$Y \sim \text{Poisson}(\exp(b_0 + b_1 * X + b_2 * \text{SES} + \text{random_effect})).$$

Now b_1 is the effect of sports after accounting for the effects of SES and so this would remove the influence of the confounder. Of course, unless we have a randomized experiment, it's possible there are other missing confounders and so we should still be careful about declaring a causal effect.

(6) Give at least three ways the `glm` function that performs non-Bayesian generalized linear models could be used in a Bayesian analysis.

(1) Initial values for the MCMC routine

(2) Empirical Bayesian priors

(3) If the sample size is much larger than the number of parameters than we could use the Bayesian CLT and approximate the posterior using the GLM estimates as the approximate posterior mean and the GLM standard error as the approximation posterior standard deviation.