ISYE 6420 Bayesian Statistics
Final Exam
Mark Pearl
December 12th, 2021

1.
a)
To handle the missing data for the chol variable, we will use the suggested distribution, $dnorm(200,\ 0.01)$

Provided is the WinBugs formulation for the problem:

```
MODEL
============================
model{
 eps <- 0.0000001
  for(i in 1:N){
   #Impute the missing data for the chol variable with suggested dnorm
   chol[i] ~ dnorm(200,0.01)
   chd[i] ~ dbern(p[i])
   logit(p[i]) <- beta[1] + beta[2] * age[i]+
           beta[3] * height[i] + beta[4] * weight[i]+ beta[5] * sdp[i] +
           beta[6] * dbp[i] + beta[7] * chol[i]  + beta[8] * cigs[i]
devres[i] <- 2*chd[i]* log(chd[i]/p[i] +eps) +
             2*(1 - chd[i])*log((1-chd[i])/(1-p[i])+eps)
 }
for(j in 1:8){
beta[j] ~ dnorm(0, 0.001)
     }
dev  <-  sum(devres[])
}

→ DATA ←

INITS
list(beta=c(0,0,0,0,0,0,0,0))
```

Figure 1: Heart WinBugs Formulation

b)
Provided are the posterior densities for the seven predictors start from the top right downwards, since the top left plot is in relation to b[1] or the intercept:

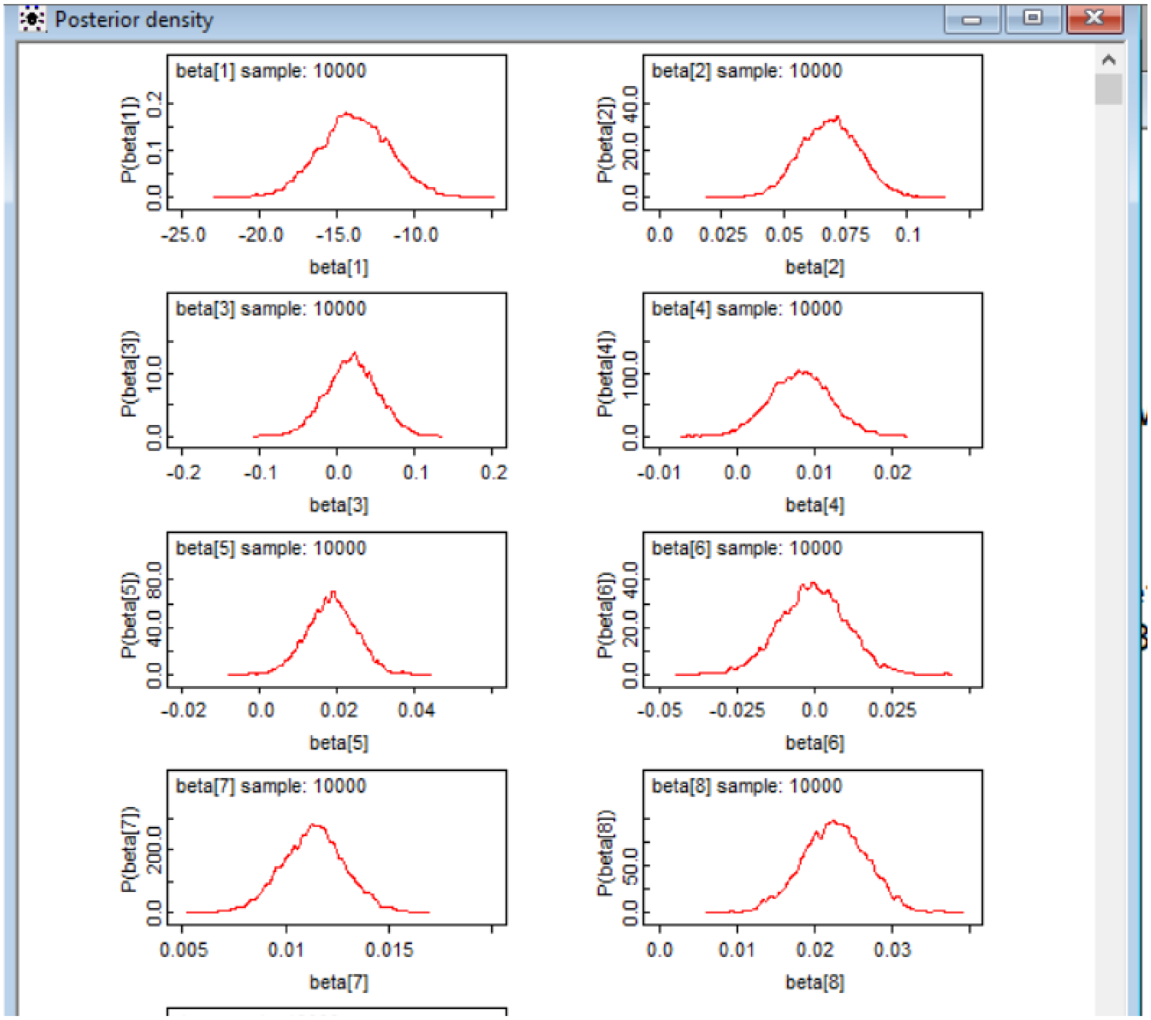


Figure 2: Density Plots Predictor Variables

Provided are the boxplots for these variables:

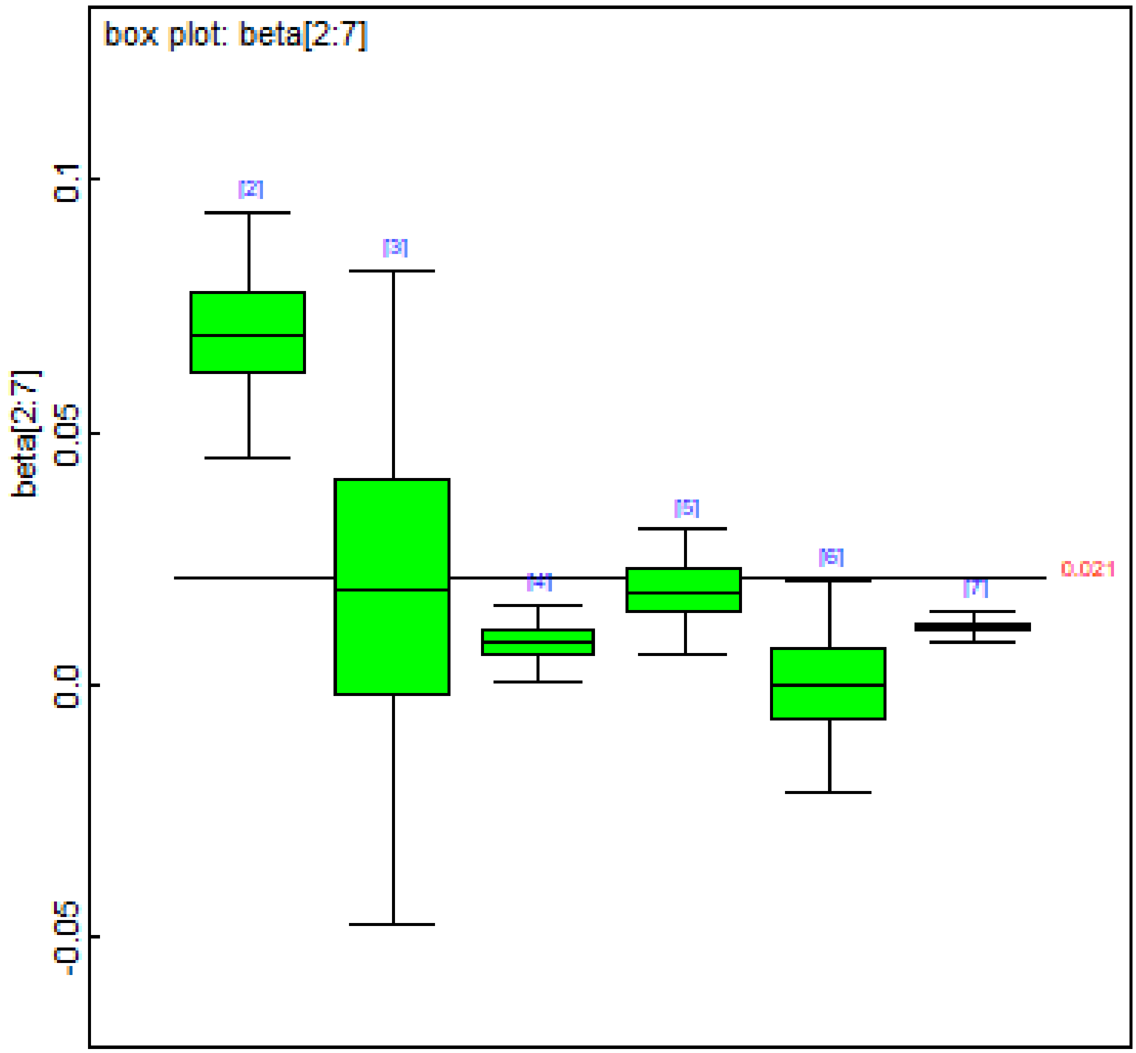


Figure 3: Boxplot Predictor Variables

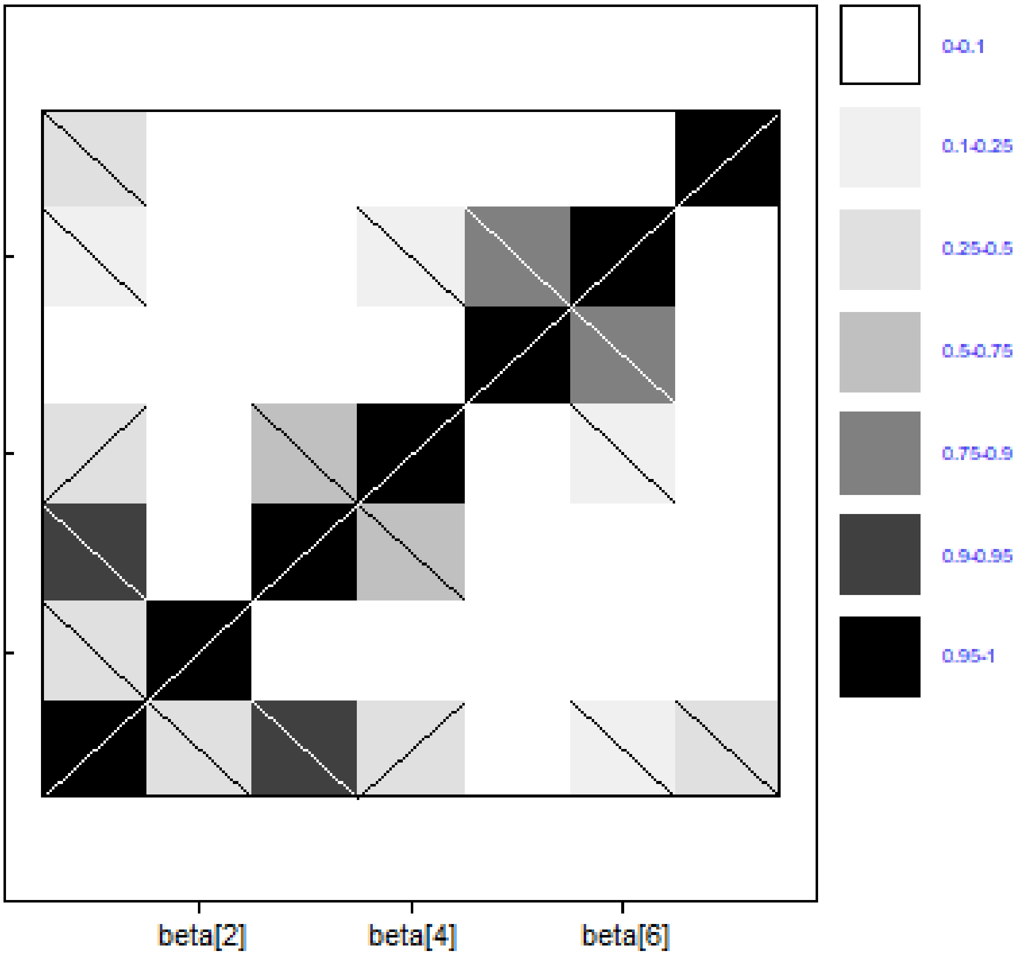


Figure 4: Correlation Plot Predictor Variables

We can also see that there's a bit of correlation exhibited between some of our variables when looking at the matrix plot.

c)
Provided is the summary output statistics for the the model:

	mean	sd	MC_error	val2.5pc	median	val97.5pc	start	sample
beta[1]	-13.86	2.29	0.03253	-18.28	-13.88	-9.367	1001	10000
beta[2]	0.0689	0.01197	1.639E-4	0.04553	0.06903	0.0924	1001	10000
beta[3]	0.01914	0.03294	4.747E-4	-0.0448	0.01941	0.08328	1001	10000
beta[4]	0.008208	0.003877	6.049E-5	6.583E-4	0.008201	0.01579	1001	10000
beta[5]	0.01853	0.006324	9.609E-5	0.00602	0.01865	0.0307	1001	10000
beta[6]	-3.034E-4	0.01078	1.723E-4	-0.02146	-2.87E-4	0.02054	1001	10000
beta[7]	0.01136	0.001531	2.256E-5	0.008346	0.01139	0.01438	1001	10000
beta[8]	0.02266	0.004204	5.995E-5	0.01421	0.02267	0.0308	1001	10000
dev	1612.0	4.068	0.06047	1606.0	1611.0	1622.0	1001	10000

Figure 5: Summary Output Heart Model

We can see that our 95% credible sets for each variable are as follows:

$age \ = \ [0.04553, \ 0.0924]$
$height \ = [-0.0448, \ 0.08328]$
$weight \ = \ [0.0006583 \ (Converted\ from\ scientific\ notation\ 6.583E-4), \ 0.01579]$
$sdp \ = \ [0.00602, \ 0.0307]$
$dbp \ = \ [-0.02146, \ 0.02054]$
$chol \ = \ [0.008346 \ , \ 0.01438]$
$cigs \ = \ [0.01421 \ , \ 0.0308]$

d)
When taking the median values to form a prediction, we're left with the following equation:
$$chd \ = \ -13.88 \ + \ 0.06903*\beta_2 + \ 0.01941*\beta_3 + \ 0.008201*\beta_4 + \ 0.01865*\beta_5 - 0.000287*\beta_6 + 0.01139*\beta_7 \ + \ 0.02267*\beta_8$$

Therefore, we will make the necessary code changes to get our p value for this prediction, based on the median values for each of the 7 predictors (calculated in excel):

OpenBUGS
File Edit Attributes Tools Info Model Inference Doodle Map Text Window Examples Manuals Help

```
 for(i in 1:N){
  #Impute the missing data for the chol variable with suggested dnorm
  chol[i] ~ dnorm(200,0.01)
  chd[i] ~ dbern(p[i])
  logit(p[i]) <- beta[1] + beta[2] * age[i]+
         beta[3] * height[i] + beta[4] * weight[i]+ beta[5] * sdp[i] +
          beta[6] * dbp[i] + beta[7] * chol[i]  + beta[8] * cigs[i]

devres[i] <- 2*chd[i]* log(chd[i]/p[i] +eps) + 2*(1 - chd[i])*log((1-
chd[i])/(1-p[i])+eps)
 }
for(j in 1:8){
beta[j] ~ dnorm(0, 0.001)
     }
dev  <-  sum(devres[])

#Prediction
median_age <- 45
median_height <- 70
median_weight <- 170
median_sdp <- 126
median_dbp <- 80
median_chol <- 223
median_cigs <- 0


logit(chd_pred) <- beta[1] + beta[2] * median_age+ beta[3] *
median_height + beta[4] * median_weight + beta[5] * median_sdp +
beta[6] * median_dbp + beta[7] * median_chol  + beta[8] * median_cigs
}


→ DATA ←

INITS
list(beta=c(0,0,0,0,0,0,0,0))
```

Figure 6: WinBugs Code Prediction Median Values

After running this we get the following result:

Node statistics

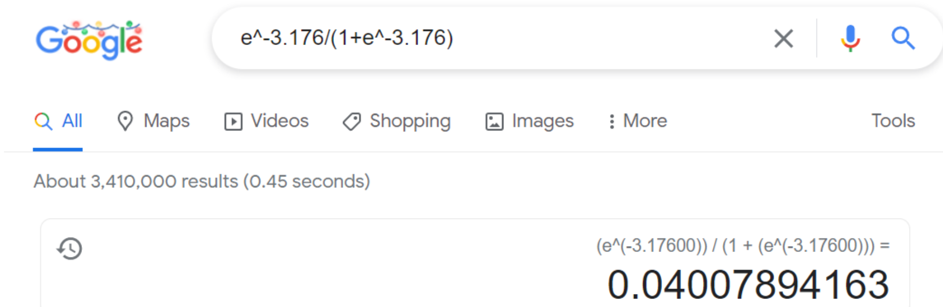
	mean	sd	MC_error	val2.5pc	median	val97.5pc	start	sample
chd_pred	0.04031	0.004395	6.621E-5	0.03235	0.04016	0.04966	1001	10000
dev	1612.0	4.068	0.06047	1606.0	1611.0	1622.0	1001	10000
deviance	102900.0	4.068	0.06047	102900.0	102900.0	102900.0	1001	10000

Figure 7: Prediction Result

We can confirm this is correct because when we remove the logit() from the result we get the -3.176. Then when plugging this into the p-value calculation where I is the result from the prediction:

$$p = \frac{e^{l}}{\left(1 + e^{l}\right)}$$

Which results in the following:



Therefore we can be confident that WinBugs is storing the result of the p-value.

Provided is our density plot for the posterior distribution of the prediction:

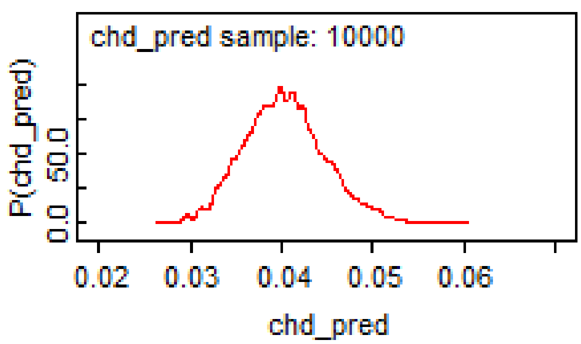


Figure 8: Density Plot chd_pred

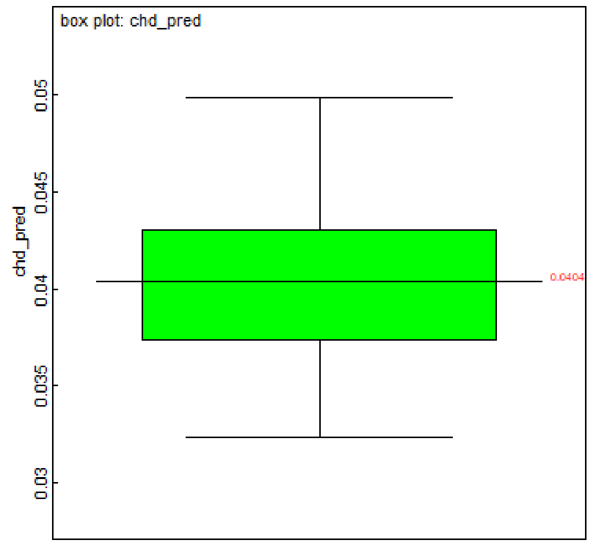


Figure 9: Boxplot chd_pred

From a mean value of 0.04031 we can conclude that this person is NOT predicted to have heart disease, assuming the default threshold of 0.5

2.
Pulp brightness. Consider the following experiment, which was performed at a pulp mill. Plant performance is based on pulp brightness as measured by a reflectance meter. Each of the four shift operators (denoted by A, B, C, and D) made five pulp handsheets from unbleached pulp. Reflectance was read for each of the handsheets using a brightness tester

Solve the problem as a Bayesian one-way ANOVA. Use STZ constraints on treatment effects.
1. Do the operators differ in making the pulp handsheets and reading their brightness? Look at the 95% credible sets for the differences between treatment effects.

We'll use the following formulation in OpenBugs to construct this problem:
We will map each operator to a given number with the following mapping:
A $\rightarrow$ 1
B $\rightarrow$ 2
C $\rightarrow$ 3
D $\rightarrow$ 4

MODEL

```
model{
for (i in 1:ntotal){
reflectance[i] ~ dnorm( mu[i], tau )
mu[i] <-  mu0 + alpha[operator[i]]
}
#alpha[1] <- 0.0;     #CR constraints
alpha[1] <- -sum( alpha[2:a] ); #STZ Constraint

mu0 ~ dnorm(0, 0.0001)
alpha[2] ~ dnorm(0, 0.0001)
alpha[3] ~ dnorm(0, 0.0001)
alpha[4] ~ dnorm(0, 0.0001)
tau ~ dgamma(0.001, 0.001)
sigma <- sqrt(1/tau)
#pairwise
  for(i in 1:3){
    for(j in i+1:4){
    adiff[i,j] <- alpha[i]-alpha[j]
    }
  }
}
```

DATA

```
list(ntotal = 20, a=4,
reflectance =
c(59.8,60,60.8,60.8,59.8,59.8,60.2,60.4,59.9,60,60.7,60.7,60.5,60
.9,60.3,61,60.8,60.6,60.5,60.5),
operator = c(1,1,1,1,1,2,2,2,2,2,3,3,3,3,3,4,4,4,4,4) )
```

INITS

```
list(mu0=0, alpha = c(NA,0,0,0), tau=1)
```

Figure 10: OpenBugs Code Problem 2

When we run our code while burning the first 1000 samples on 100,000 iterations we get the following result:

Node statistics

	mean	sd	MC_error	val2.5pc	median	val97.5pc	start	sample
adiff[1,2]	0.1804	0.2208	7.127E-4	-0.2568	0.1798	0.6173	1001	100000
adiff[1,3]	-0.3793	0.2205	6.888E-4	-0.816	-0.3792	0.05947	1001	100000
adiff[1,4]	-0.4398	0.221	6.914E-4	-0.8776	-0.4394	-0.002637	1001	100000
adiff[2,3]	-0.5597	0.2205	9.809E-4	-0.9958	-0.5604	-0.1247	1001	100000
adiff[2,4]	-0.6202	0.2202	9.824E-4	-1.057	-0.6204	-0.1823	1001	100000
adiff[3,4]	-0.06052	0.2212	9.542E-4	-0.5	-0.0598	0.3787	1001	100000
alpha[1]	-0.1597	0.1353	3.106E-4	-0.4286	-0.1597	0.1096	1001	100000
alpha[2]	-0.3401	0.1349	5.832E-4	-0.6077	-0.3402	-0.07287	1001	100000
alpha[3]	0.2196	0.1352	5.64E-4	-0.0471	0.2199	0.4884	1001	100000
alpha[4]	0.2801	0.1353	5.655E-4	0.01414	0.28	0.5471	1001	100000
mu0	60.4	0.07784	2.435E-4	60.25	60.4	60.55	1001	100000
sigma	0.3425	0.06506	2.665E-4	0.2429	0.3331	0.4963	1001	100000

Figure 11: Reflectance Summary Output

When analyzing the output, we can confirm that some of the operators differ in making pulp handsheets as some of the credible sets for our pairwise differences do not contain 0. When we look at the combinations:

A and D seem to differ since the credible set ranges from -0.8776 to -0.002637

Meaning $\alpha_a < \alpha_d$

B and C seem to differ since the credible set ranges from -0.9958 to -0.1247

Meaning $\alpha_b < \alpha_c$

B and D seem to differ since the credible set ranges from -1.057 to -0.1823

Meaning $\alpha_b < \alpha_d$

Since the credible sets are all negative, H_0 being the assumption that the treatment means are equal is rejected.

b)

Find the 95% credible set for the contrast μ1 − μ2 − μ3 + μ4, where μ1, μ2, μ3, and μ4 are the mean pulp brightness for the operators A, B, C, and D, respectively.

When we reformulate our WinBugs code, we initialize this case to a fixed index of $adiff[7,7]$. This results in the following code change:

tester.

MODEL

```
model{
for (i in 1:ntotal){
reflectance[i] ~ dnorm( mu[i], tau )
mu[i] <-  mu0 + alpha[operator[i]]
}
#alpha[1] <- 0.0;     #CR constraints
alpha[1] <- -sum( alpha[2:a] ); #STZ Constraint

mu0 ~ dnorm(0, 0.0001)
alpha[2] ~ dnorm(0, 0.0001)
alpha[3] ~ dnorm(0, 0.0001)
alpha[4] ~ dnorm(0, 0.0001)
tau ~ dgamma(0.001, 0.001)
sigma <- sqrt(1/tau)
#pairwise
  for(i in 1:3){
    for(j in i+1:4){
    adiff[i,j] <- alpha[i]-alpha[j]
    }
  }
adiff[7,7] <- (alpha[1]-alpha[2])-(alpha[3]-alpha[4])

}
```

When we run we get the same output but with the additional adiff entry added:

Node statistics

	mean	sd	MC_error	val2.5pc	median	val97.5pc	start	sample
adiff[1,2]	0.1804	0.2208	7.127E-4	-0.2568	0.1798	0.6173	1001	100000
adiff[1,3]	-0.3793	0.2205	6.888E-4	-0.816	-0.3792	0.05947	1001	100000
adiff[1,4]	-0.4398	0.221	6.914E-4	-0.8776	-0.4394	-0.002637	1001	100000
adiff[2,3]	-0.5597	0.2205	9.809E-4	-0.9958	-0.5604	-0.1247	1001	100000
adiff[2,4]	-0.6202	0.2202	9.824E-4	-1.057	-0.6204	-0.1823	1001	100000
adiff[3,4]	-0.06052	0.2212	9.542E-4	-0.5	-0.0598	0.3787	1001	100000
adiff[7,7]	0.2409	0.3119	0.001191	-0.3752	0.2413	0.8573	1001	100000
alpha[1]	-0.1597	0.1353	3.106E-4	-0.4286	-0.1597	0.1096	1001	100000
alpha[2]	-0.3401	0.1349	5.832E-4	-0.6077	-0.3402	-0.07287	1001	100000
alpha[3]	0.2196	0.1352	5.64E-4	-0.0471	0.2199	0.4884	1001	100000
alpha[4]	0.2801	0.1353	5.655E-4	0.01414	0.28	0.5471	1001	100000
mu0	60.4	0.07784	2.435E-4	60.25	60.4	60.55	1001	100000
sigma	0.3425	0.06506	2.665E-4	0.2429	0.3331	0.4963	1001	100000

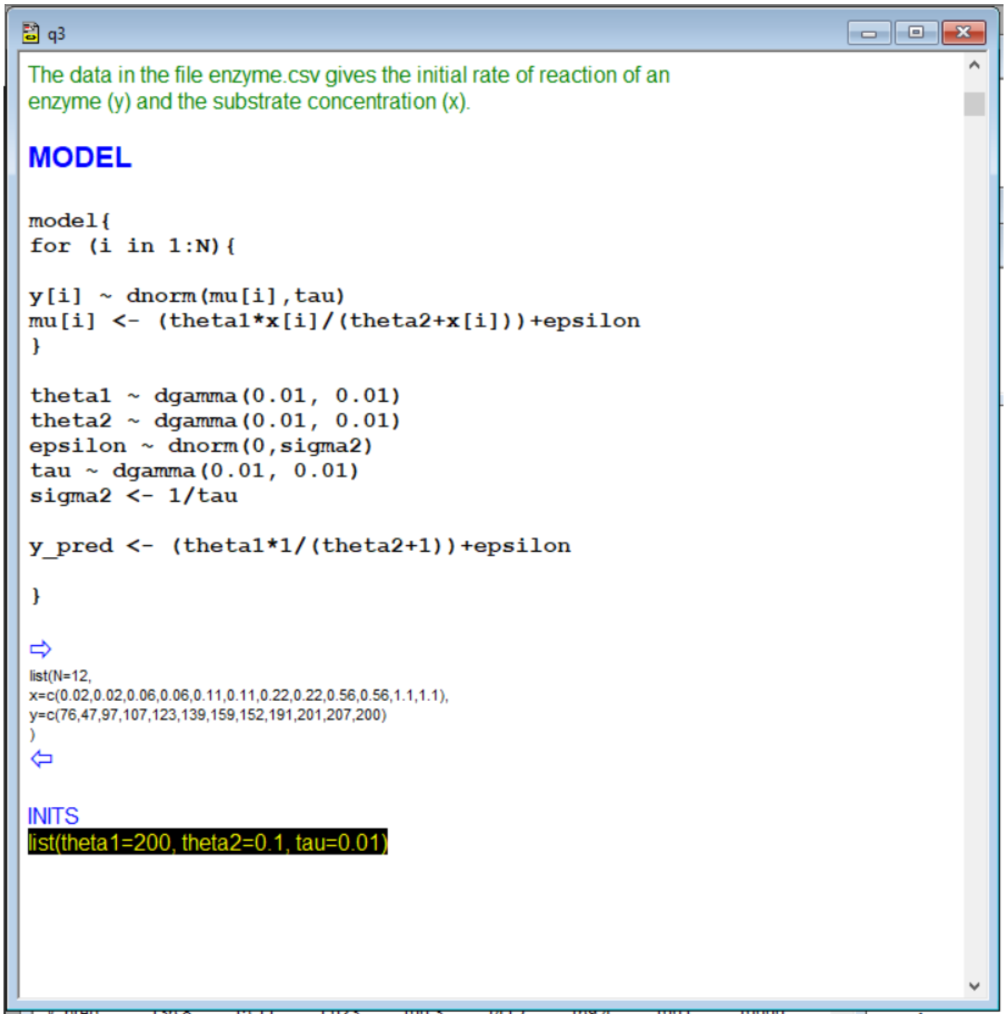
Figure 12: Summary Output Ctd

Therefore our output for the credible set is $[-0.3752,\ 0.8573]$

3.

a)

Provided is the winbugs formulation for the problem:



q3

The data in the file enzyme.csv gives the initial rate of reaction of an enzyme (y) and the substrate concentration (x).

MODEL

```
model{
for (i in 1:N){

y[i] ~ dnorm(mu[i],tau)
mu[i] <- (theta1*x[i]/(theta2+x[i]))+epsilon
}

theta1 ~ dgamma(0.01, 0.01)
theta2 ~ dgamma(0.01, 0.01)
epsilon ~ dnorm(0,sigma2)
tau ~ dgamma(0.01, 0.01)
sigma2 <- 1/tau

y_pred <- (theta1*1/(theta2+1))+epsilon

}
```

```
list(N=12,
x=c(0.02,0.02,0.06,0.06,0.11,0.11,0.22,0.22,0.56,0.56,1.1,1.1),
y=c(76,47,97,107,123,139,159,152,191,201,207,200)
)
```

INITS

```
list(theta1=200, theta2=0.1, tau=0.01)
```

Figure 13: WinBugs Formulation Enzymes

This produced the following plots for our densities:

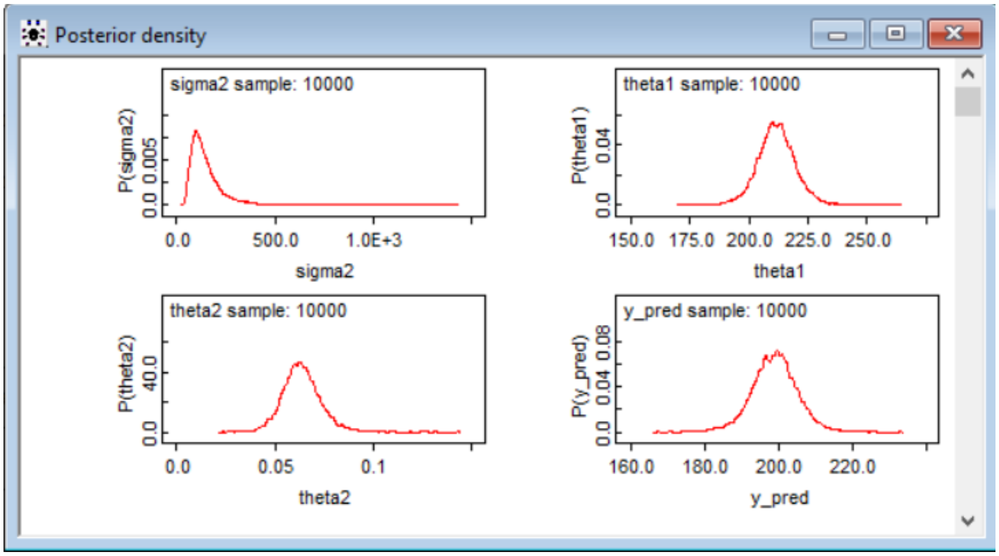


Figure 14: Marginal Densities for each variable

b)

Running this model gives us the following summary output:

Node statistics

	mean	sd	MC_error	val2.5pc	median	val97.5pc	start	sample
sigma2	148.6	85.54	1.196	57.57	127.2	361.9	1001	10000
theta1	211.6	7.92	0.18	195.7	211.6	227.7	1001	10000
theta2	0.06361	0.009662	2.195E-4	0.04588	0.06307	0.08446	1001	10000
y_pred	198.9	6.147	0.1296	186.3	199.0	211.2	1001	10000

Figure 15: Enzymes Summary Output

Therefore our credible sets for each 3 variables are:

$sigma2\ (sigma\ squared) \ = \ [55.57,\ 361.9]$

$\theta_1 \ = [195.7,\ 227.7]$

$\theta_2 \ = \ [0.04588,\ 0.08446]$

c)

When we plot the density for where x = 1, we get the following:

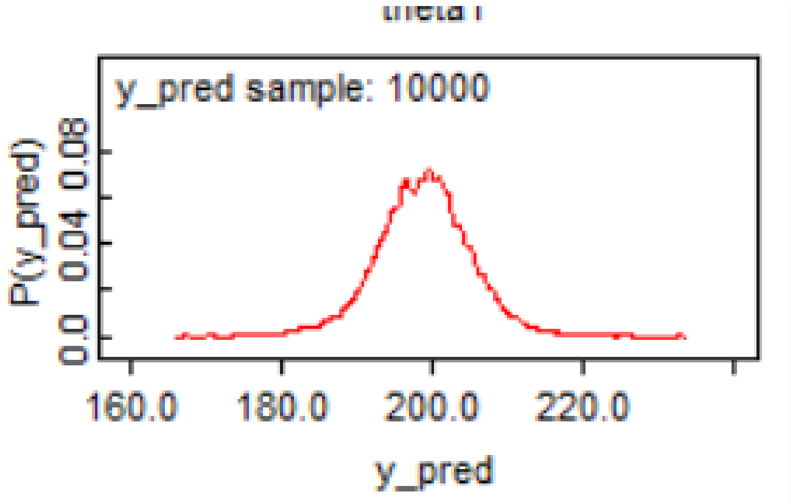


Figure 16: Prediction Density Plot

Which also produces the following credible set:
$y_{pred} \ = \ [186.3, \ 211.2]$