DSA5403: FINAL

Instructor: Dr. Wayne Stewart

Monday May 9, 7:00 am

About the exam and Instructions:

- EXAM LENGTH: 24 hours
- Please answer all questions
- This exam will cover chapters 1-13, 15,18 and 19 of "Doing Bayesian Data Analysis" second edition by JK and the related course material for DSA 5403 including labs and quiz questions
- All models and R code used to answer questions must be within the RMD document and finally compiled to make html
- Use the blank RMD document provided and place all R and Jags code in code chunks so that answers will be made through the document.
- You will upload the RMD and HTML documents and I will check your code by running the rmd.

Question 1:

A researcher wishes to investigate whether a coin is biased. She throws the coin 10 times and obtains 4 heads (successes). Suppose θ is the probability of a head. That is $n = 10, x = 4$. The researcher wishes to use a Beta prior to express her prior belief about the distribution of θ.

1a) Derive the general formula for the posterior $p(\theta|x)$

Part of this might be already done

The posterior should be in terms of $n, x, \theta, \alpha, \beta$.

You may need the following formulae:

$$p(X = x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

$$p(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$$

1b) If $\alpha = \beta = 1$ what prior distribution does this correspond to?

You will need to correctly name the distribution. This prior was not chosen to make the posterior.

1c) Suppose she uses $\alpha = \beta = 5$ for the prior what is the posterior distribution?

You will need to give the name of the distribution and the values of the parameters.

1d) If a further experiment is made after the first using the *same* coin, this time with $n = 20, x = 12$, what prior should the researcher use in the absence of any other information than what is given in this problem?

You must give the name of the distribution and its parameter values.

1e) Find the posterior after the second sequential experiment.

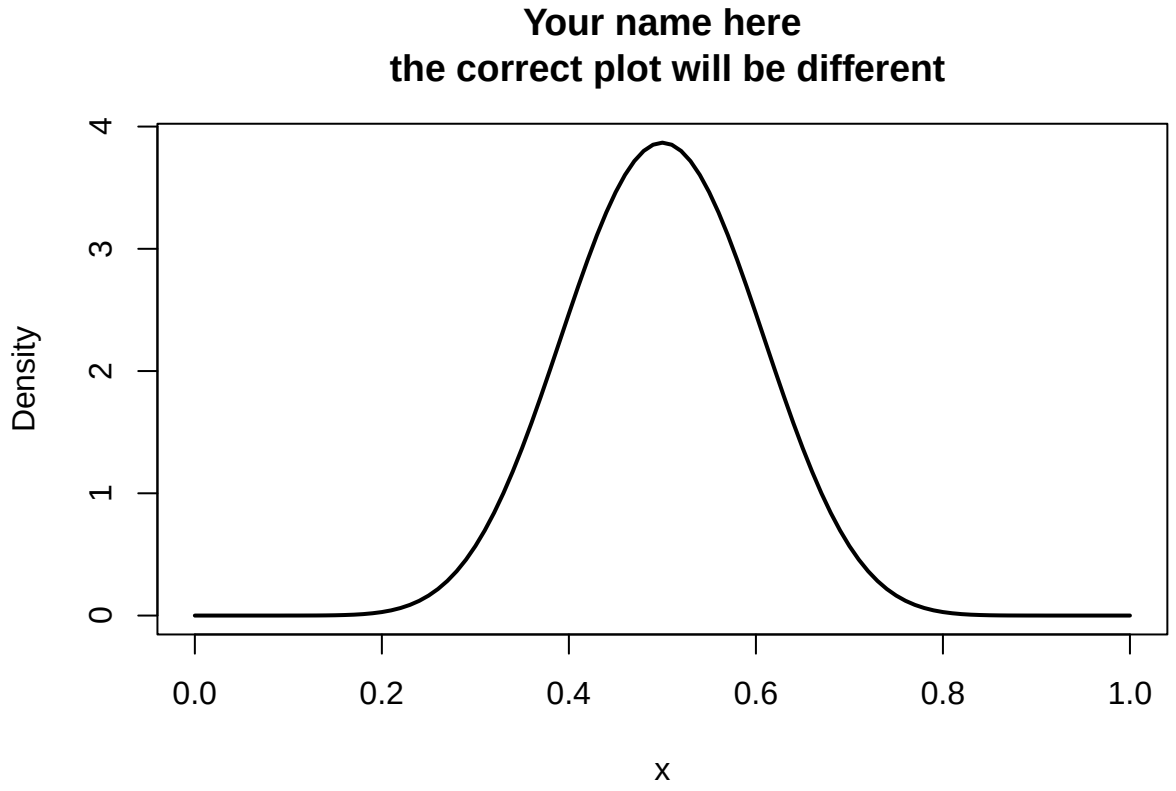
You will need to show your working and derive the result analytically.

1f) The researcher now wishes to plot the posterior (that which is formed from the two experiments). Fill in the gaps of the code so that the correct plot is created.

1f i) A=

1f ii) B=

```
curve(dbeta(x,A,B), xlim = c(0,1), lwd=2, ylab ="Density",main="Your name here")
```



Question 2:

Suppose $X \sim Beta(\alpha, \beta)$

Then the following comes from R help:

```
dbeta(x, shape1, shape2, ncp = 0, log = FALSE)
pbeta(q, shape1, shape2, ncp = 0, lower.tail = TRUE, log.p = FALSE)
qbeta(p, shape1, shape2, ncp = 0, lower.tail = TRUE, log.p = FALSE)
rbeta(n, shape1, shape2, ncp = 0)
```

2 Write down the code that answers the following - you *MUST* use default options

2a) $P(X > 0.7|\alpha = 2, \beta = 3)$

Probabilities, but cannot use 1 - etc.

2b) $P(X < 0.2|\alpha = 1, \beta = 1)$

2c) Find the value of X with lower tail probability 0.04. $X \sim Beta(5, 4)$

2d) Find the equal tail interval that contains 95% of the distribution, $X \sim Beta(4, 8)$

2e) Generate a random sample of X values of size 20 with $X \sim Beta(5, 7)$

Question 3 Mostly Easy - interpretation - some from the book

In performing Bayesian analyses we often need to interpret output of posteriors. The following is taken from Figure 8.4 in the book.

```
plotPost( codaSamples[,"theta"] , main="theta" , xlab=bquote(theta) ,
          cenTend="median" , compVal=0.5 , ROPE=c(0.45,0.55) , credMass=0.90 )
```

Notice that there were four additional arguments specified. The `cenTend` argument specifies which measure of central tendency will be displayed; the options are `"mode"`,

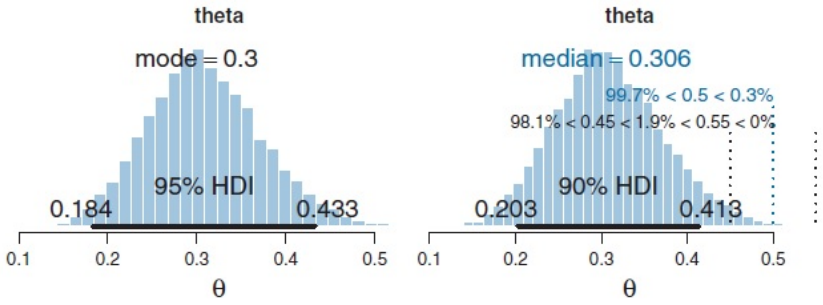


Figure 8.4 Posterior distribution based on output from JAGS, plotted twice using different options in the `plotPost` function.

Please answer the following from the pictures

3a) Give the 95% HDI for θ the probability of a success.

3b) What is the posterior probability that $\theta < 0.5$

You will need to read the answer off the appropriate picture.

3c) If $H_0 : \theta = 0.5$ should we reject it (the NULL), yes/no?

Give the answer and explain why you decided this.

3d) The posterior probability that θ lies in $(0.203, 0.413)$ is 0.90. True or False?

You will need to choose and explain your answer.

3e) What does ROPE stand for?

You will need to say what each letter means and also explain how the ROPE is used in a Bayesian analysis.

Question 4: Easy - interpretation

In Lab 12 and Ass 4 you were asked to analyze the Titanic data set using a Bayesian model.

The data was manipulated and formatted in the following way:

```
clglm = glm(survived ~ sex + age + sex:age,family = "binomial", data = Titanicp)

mat1=model.matrix(clglm)
mat2=model.frame(clglm)

y = with(mat2, ifelse(survived == "survived", 1,0))

dataList=list(y = y, x = mat1[, "age"],sexm = mat1[,"sexmale"], sexmx = mat1[,"sexmale:age"] , n = leng
```

The following is part of the Jags code needed to run the model:

```
library(rjags)

#Define the model:
modelString = "
model{
  for(i in 1:n){
  y[i] ~ dbin(theta[i], 1)
   eta[i]<- beta[1] + beta[2]*x[i] + beta[3]*sexm[i] + beta[4]*sexmx[i]
  logit(theta[i]) <- eta[i]
  }

for(j in 1:4){
```

```
  beta[j] ~ dnorm(0,1.0E-3)
}
}
"
writeLines( modelString , con="TEMPmodel.txt" )

initsList = list(beta = c(0.5,0.02,-1.15,-0.05))

# Run the chains:
jagsModel = jags.model( file="TEMPmodel.txt" , data=dataList , inits=initsList ,
                        n.chains=3 , n.adapt=500 )

update( jagsModel , n.iter=500 )
codaSamples = coda.samples( jagsModel , variable.names=c("beta"),
                            n.iter=3330 )
save( codaSamples , file=paste0("lab12","Mcmc.Rdata") )



library(ggmcmc)
s = ggs(codaSamples)
d=ggs_density(s)

print(d)

cr =  ggs_crosscorrelation(s)
print(cr)

summary(codaSamples)
```

From the code above answer the following questions

4a) What is the link function used here?

You will need to not only name it as a function but also define it in terms of `theta` the probability of a success.

4b) What values does `y` take?

That is when used in the model what numerical values will replace the `y` variable?

4c) Are the priors used for the `beta` components high impact?

Say yes or no and give explanation about what high/low impact means

4d) Below is some output: please review and answer the fiollowing

4d 1) What is the point estimate for β_3 ?

4d 2) What is the 95% BCI for β_1

```
summary(codaSamples)
##
## Iterations = 1001:4330
## Thinning interval = 1
## Number of chains = 3
## Sample size per chain = 3330
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##             Mean       SD  Naive SE Time-series SE
## beta[1]  0.31744 0.151638 1.517e-03      0.0109514
## beta[2]  0.01326 0.004944 4.947e-05      0.0003840
## beta[3] -0.75905 0.199376 1.995e-03      0.0146046
## beta[4] -0.02623 0.006402 6.406e-05      0.0004753
##
## 2. Quantiles for each variable:
##
##              2.5%       25%      50%      75%   97.5%
## beta[1]  0.025778  0.212207  0.31721  0.41899  0.6118
## beta[2]  0.003728  0.009919  0.01317  0.01667  0.0228
## beta[3] -1.141633 -0.896826 -0.76158 -0.62038 -0.3779
## beta[4] -0.038773 -0.030569 -0.02613 -0.02188 -0.0136
```

Question 5: Unkown - will need to research

The exponential family can be defined as any density of the following form:

$$f(y|\theta,\phi) = exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right)$$

5a) Show using algebra and the formula for the exponential family that $E(Y) = b^{'}(\theta)$

5b) Show using algebra and the formula for the exponential family that $V(Y) = b^{''}(\theta)a(\phi)$

Suppose $y \sim Binomial(n,p)$

where:

$$p(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}$$

where

$$\binom{n}{y} = \frac{n!}{y!(n-y)!}$$

5c) Show that the Binomial distribution belongs to the exponential family

All working must be shown!

5d) Show that $E(Y) = np$ using the above results.

This must be done with clarity!! Algebra must be used that clearly demonstrates this!!

5e) Show that $V(Y) = np(1-p)$ using the above results.

This must be done with clarity!! Algebra must be used that clearly demonstrates this!!

Question 6: Some done before - check lab and Assn (possibly prior exam???)

Bayes' theorem can be written as:

$$p(\theta|x) = \frac{p(\theta) f(x|\theta)}{p(x)}$$

If the prior is a Beta :

$$p(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}$$

Derive the formula for the evidence by substituting the expression for the posterior (see qu1a) and rearranging and simplifying for $p(x)$

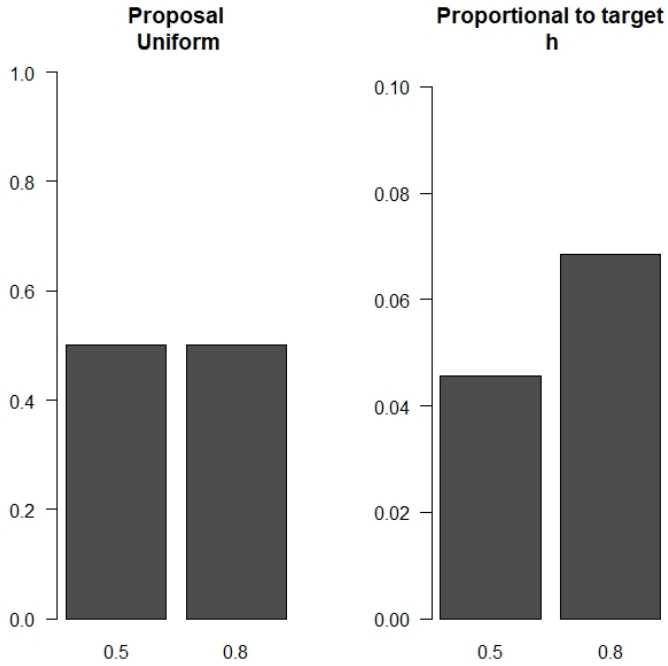
Question 7: Midterm and lab - not all the way but can backtrace

A coin was tossed 10 times with the result that there were 6 heads. It was thought that the probability of a head could only be 0.5 or 0.8. The following Bayes' box was generated with $h = prior * lik$:

```
##        theta     prior        lik          h post
## 1        0.5 0.2225952 0.20507812 0.04564941  0.4
## 2        0.8 0.7774048 0.08808038 0.06847411  0.6
## Totals    NA 1.0000000         NA 0.11412352  1.0
```

The Bayes' box with a 2 state uniform proposal was used to generate the following MCMC.

A coin was used as a proposal and a die supplied the acceptance values.



The proposal and posterior are states, where state 1 corresonds to theta = 0.5 and state 2 corresponds to theta = 0.8. E is the acceptance set and $E2 = \{3, 4, 5, 6\}$ for the die value.

Answer the questions by examining the output below!

```
   proposal alpha  E    dice post
1  "1"      "1"    "E1" "2"  "1"
2  "2"      "1"    "E1" "6"  "A" <--
3  "2"      "1"    "E1" "6"  "2"
4  "1"      "0.67" "E2" "2"  "2"
5  "2"      "1"    "E1" "6"  "2"
6  "1"      "0.67" "E2" "4"  "1"
7  "1"      "1"    "E1" "3"  "1"
8  "2"      "1"    "E1" "2"  "2"
9  "2"      "1"    "E1" "2"  "2"
10 "2"      "1"    "E1" "1"  "2"
11 "1"      "B"    "E2" "4"  "1" <--
12 "1"      "1"    "E1" "3"  "1"
13 "2"      "1"    "E1" "6"  "2"
14 "1"      "0.67" "E2" "3"  "1"
15 "1"      "1"    "E1" "2"  "1"
16 "1"      "1"    "E1" "5"  "1"
17 "2"      "1"    "E1" "5"  "2"
18 "1"      "0.67" "E2" "6"  "C" <--
19 "1"      "1"    "E1" "3"  "1"
20 "1"      "1"    "E1" "1"  "1"
```

7a) Find the value of A

7b) Find the value of B

7c) Find the value of C

7d) Write down the formula for the acceptance probability α_{ij}

Question 8 Mainly from prior exam or lab

Suppose a researcher gathered data that suited a simple linear regression (SLR). The data came as a data frame of x and y values with a total of n rows. Then the appropriate model would be

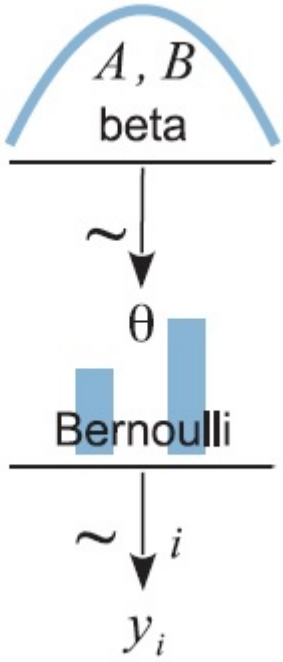
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

and

$$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

8a) Either sketch a Directed Acyclic Graph (also known as a doodle) or sketch a model (as in the text book) diagram below for the SLR model.

This means using ellipses and a rectangle with arrows named appropriately or else a model diagram like (but not the same as) Figure 2



8b) Write down the jags code for the model using uniform's on the sigma's

8c) What node would be called `logical`? ?

8d) What is the linear predictor in this model?

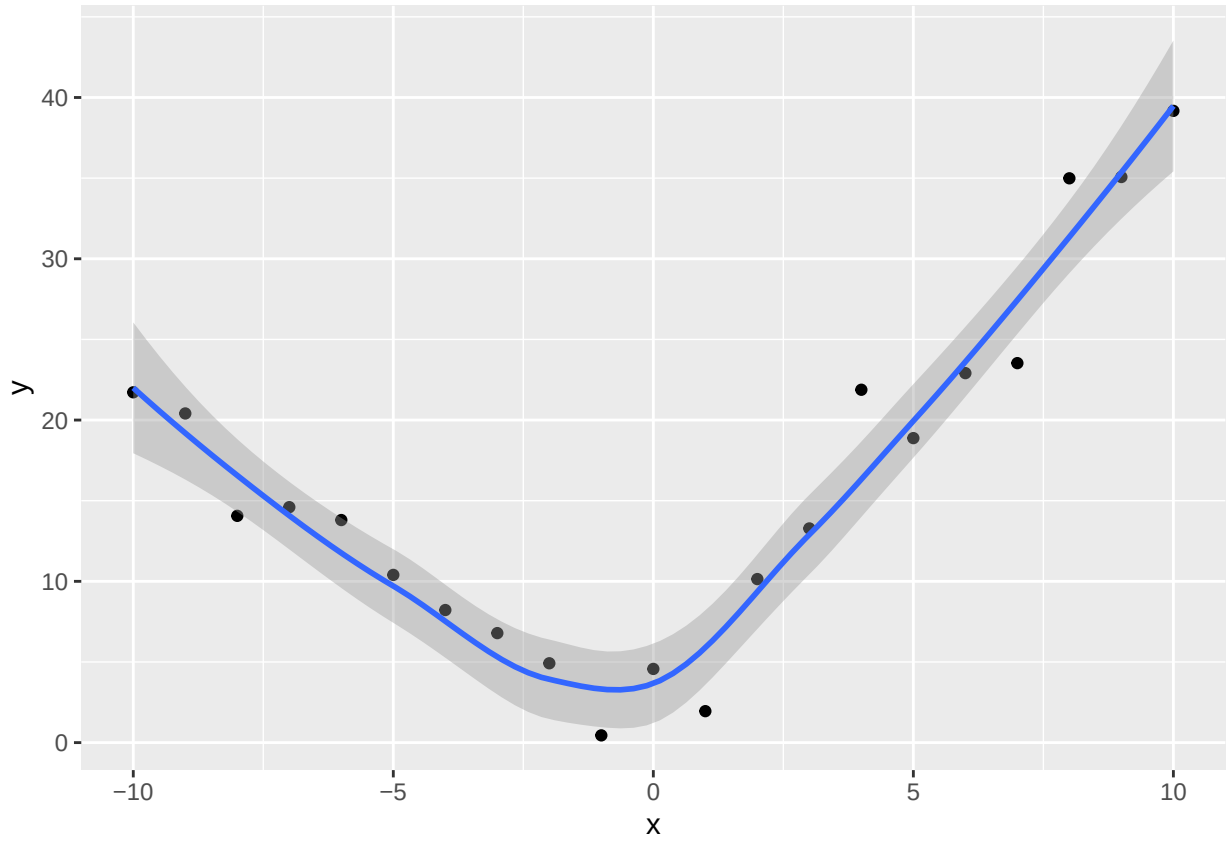
Question 9 Not same but similar to the piecewise lab

The following changepoint data is to be plotted and then a Bayesian model used to estimate θ the x value where a change in slope is made and the other parameters needed to characterize the model.

```
dataList = list(x = c(-10, -9, -8, -7, -6, -5, -4, -3, -2, -1, 0, 1, 2,
3, 4, 5, 6, 7, 8, 9, 10), y = c(21.72, 20.41, 14.06, 14.6, 13.8,
10.4, 8.22, 6.79, 4.92, 0.45, 4.57, 1.95, 10.14, 13.28, 21.88,
18.88, 22.91, 23.53, 34.99, 35.07, 39.17), N = 21)
```

9a) Create the following ggplot

```
## `geom_smooth()` using formula 'y ~ x'
```



9b) Write JAGS code using low impact priors to find an interval and point estimate for θ you will need to estimate the other parameters as well.

9b) i)

Write the JAGS code

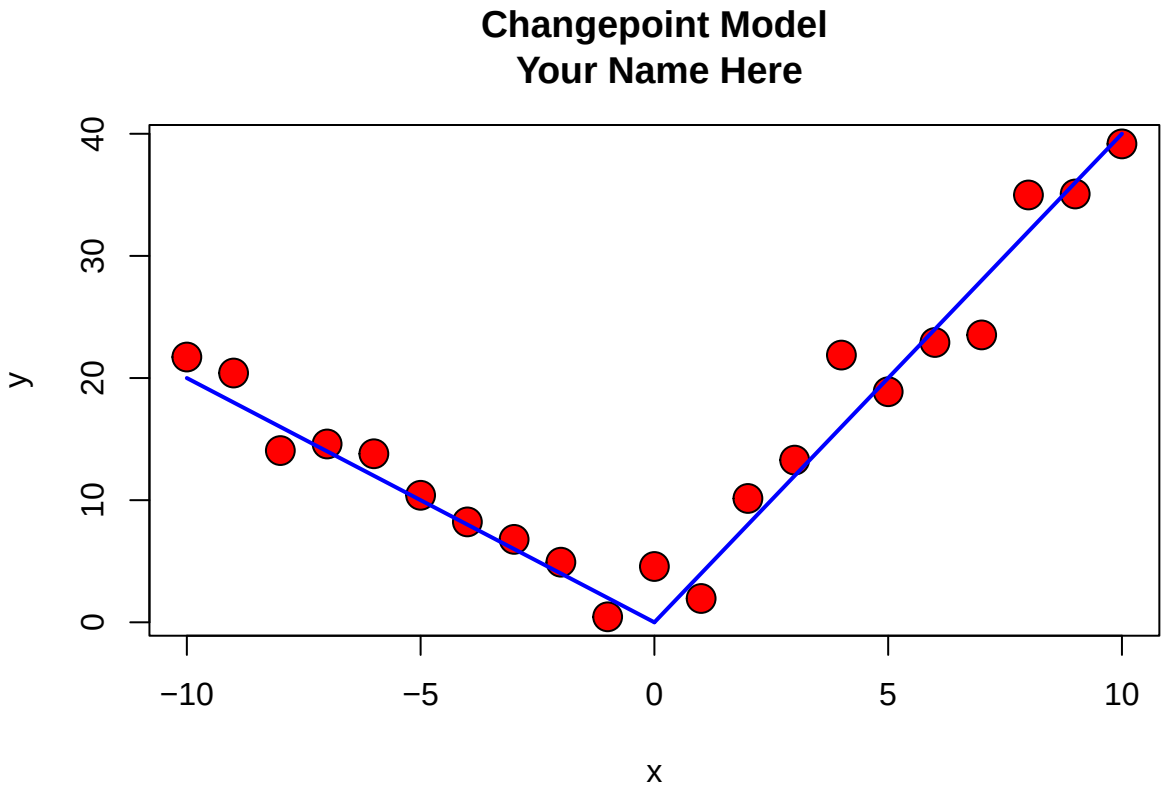
9b) ii)

Summarize θ

Interpret the value of theta using both the point estimate and 95% BCI

9b) iii)

Now plot the estimating lines using the parameters you have estimated. All code must be in this document.



Question 10

Pretty basic - likely using centering makes the difference

Make two SLR Bayesian models to estimate the intercept β_0 in each case.

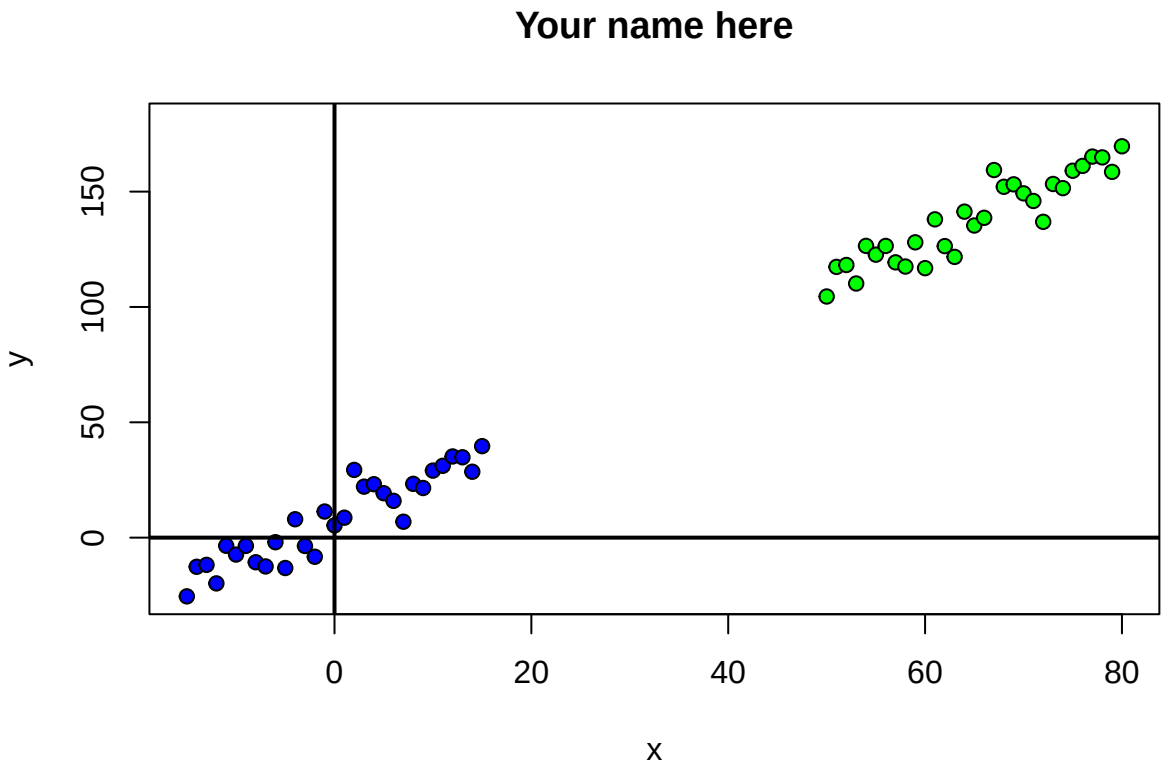
Data:

```r
x=50:80
set.seed(24)
y=10+2*x+rnorm(31,0,10)

xx=-15:15
set.seed(24)
yy=10+2*xx+rnorm(31,0,10)
```

Question 10a)

Plot the data



Question 10b)

Find point and 95% interval estimates for the β's in each model by making two JAGS models

Question 10b) i)

Jags model 1 here

Question 10b) ii)

Jags model 2 here

Question 10c)

Give two posterior β_0 density plots. One for the left most data and the other for the right most data.

Question 10d)

Explain why the intervals for β_0 are so different. Make reference to the density plots also.

Question 10e)

Use the function `lm`() and make classical 95% confidence interval estimates for the β values for both sets of data – compare within the classical setting and then compare the Bayesian to the Classical estimates.