Bayesian Data Science: Assignment 4

Instructions:

- Please do ***all*** questions and parts.
- Please show all relevant working and use R and Jags for all statistical programming and analysis. This assignment will cover part 3 of the course: Chapters 15 and following "*Doing Bayesian Data Analysis*", second edition by JK.
- This assignment will be graded on 100 points. This third part of the course is where we introduce the glm (generalized linear model), it

… applies the Bayesian methods to realistic data. The applications are organized around the type of data being analyzed, and the type of measurements that are used to explain or predict the data. Different types of measurement scales require different types of mathematical models, but otherwise the underlying concepts are always the same.

- Use "R Mark down" to construct your assignment answers. Use appropriate Latex formulae (https://en.wikibooks.org/wiki/LaTeX/Mathematics) and r code chunks.
  - Please submit your Rmd document, html and pdf knit documents.
  - See http://rmarkdown.rstudio.com/index.html for help with rmarkdown
  - Make sure all of the files pertaining to the assignment are in the same directory.
  - I will run your Rmd document and check that your code works.
  - The functions you make must have arguments suitable to solve the problem and produce the desired output. ***It is up to YOU to determine what arguments are needed***.
  - You may use base R functions and/or other packages like ggplot2
  - ***Make sure your name is on the titles of all plots you make.***
  - ***You may consult google, youtube etc for help on R, ggplot2, JAGS etc***
  - All plots should be large (at least ¼ a page)
- This assignment assumes that you have already installed the following software (all of which are free)
  - R
  - R Studio
  - Latex (best to have a full distribution)
  - Jags
- Your finished assignment should be readable and intelligible.
- ***You are expected to do ALL questions and parts!!!***
  - ***For those who did not do them all please give a list of all questions and parts you DID NOT DO!! This will be the last thing you do before uploading the documents.***
  - ***If you did all questions and parts please say: "All questions and parts done completely"***
  - ***Put this statement at the top of the document.***
- This assignment will start with help and guidance and end requiring more of your own application of knowledge with less help from me.
- ***The last question is on the Titanic data set – this is the culmination of the course and I will be looking for evidence that you can perform Bayesian statistics by yourself and be creative.***

Questions:

1. ***Soft Drink Delivery Times***. The following example deals with the quality of the delivery system network of a soft drink company; see example 4.1 in Montgomery & Peck (1992). In this problem, interest lies in the estimation of the required time needed by each employee to refill an automatic vending machines owned and served by the company. For this reason, a small quality assurance study was set-up by an industrial engineer of the company. As ***the response variable*** he considered the ***total service time*** (measured in minutes) of each machine including its stocking with beverage products and any required maintenance or housekeeping. After examining the problem, the industrial engineer recommended two important variables which affect ***delivery time:***
    1. the number of cases of stocked products and
    2. the distance walked by the employee (measured in feet).

   A dataset of 25 observations was finally collected. Below is part of a Bayesian model ***with possible errors and missing code***. You can use this as a help in making your own model for jags.

```
model{
# model’s likelihood
for (i in 1:n){
time[i] ~ dnorm( mu[i], tau ) # stochastic componenent
# link and linear predictor
mu[i] <- beta0 + beta1 * cases[i] + …
}
# prior distributions
tau ~ dgamma( 0.01, 0.01 )
…
…
beta2 ~ dnorm( 0.0, 1.0E-4)
…
…
}
INITS
list( tau=1, beta0=1, beta1=0, beta2=0 )
DATA (LIST)
list( n=25,
time = c(16.68, 11.5, 12.03, 14.88, 13.75, 18.11, 8, 17.83,
79.24, 21.5, 40.33, 21, 13.5, 19.75, 24, 29, 15.35,
19, 9.5, 35.1, 17.9, 52.32, 18.75, 19.83, 10.75),
distance = c(560, 220, 340, 80, 150, 330, 110, 210, 1460,
605, 688, 215, 255, 462, 448, 776, 200, 132,
36, 770, 140, 810, 450, 635, 150),
cases = c( 7, 3, 3, 4, 6, 7, 2, 7, 30, 5, 16, 10, 4, 6, 9,
10, 6, 7, 3, 17, 10, 26, 9, 8, 4) )
```

a) What is the mathematical model used here in terms of the GLM? Write out the mathematics using Latex

   i. What is the ***Link*** in this case?

ii. What is the ***linear predictor*** in this case?
iii. What is the ***distribution*** of the error in this case?

b) Plot the data as boxplots using ggplot (you will need to reformat the data as a data frame, `data.frame()`)

c) Make density plots of all stochastic nodes using the `ggmcmc` package. You should run the model for 10000 iterations and throw away the first 1000 as a burn in. Use three chains.

d) Give diagnostic plots of the MCMC for all primary stochastic nodes.

e) Make posterior histograms of beta0, beta1 and beta2 using JK's code with a suitable rope around 0.

f) Give a summary of all the posterior parameter estimates.

g) Give all Bayesian point estimates of parameters.

h) Write down the formula (using Latex) for the mean service (or delivery) time as predicted by the model

i) Using the above expression for the mean service time and summary information from the posterior find:

i. For each additional case stocked by the employee
   a. how much delivery time will be required on average (point estimate)?
   b. how much delivery time will be required on average (interval estimate) and with what posterior probability?

ii. For every increase of walking distance by 100 feet
   a. what delivery time will be required on average (point estimate)?
   b. what delivery time will be required on average (interval estimate) and with what posterior probability?

j) The engineer wished to find a typical or representative delivery route. He suggested the following code chunk. Complete the code by supplying the missing function

```
typical.y <- beta0 + beta1 * mean(cases[]) + beta2 * …
```

k) Add the typical.y to the JAGS model and re-run this time including `typical.y` as a monitored node. Give a full interpretation of its posterior distribution.

l) $R_B^2 = 1 - \frac{\sigma^2}{s_Y^2}$. s_Y^2 is the sample variance of Y and σ^2 is the variance of the random variable Y in the model. R_B^2 is the reduction in uncertainty of the response variable Y through the model using explanatory variables x. (Bayesian analog of classical R_{adj}^2.)

i. The following code is incomplete – give the rest of it as would be needed to calculate R_B^2

```
# definition of sigma
s2<-1/tau
s <-sqrt(s2)
# calculation of the sample variance
for (i in 1:n){ c.time[i]<-time[i]-mean(time[]) }
sy2 <- inprod( c.time[], c.time[] )/(n-1)
# calculation of Bayesian version of adj R squared
R2B <- …
```

ii. Include the code in the model and re-run after monitoring `R2B` – give a full interpretation of the `R2B` output.

iii. Include code that will calculate $P(\beta_2 > 0.01|D)$ where D is the data. Hint: Use `step()`

iv. Find the point estimate for the above probability.

2. Dobson (1983) analyses binary dose-response data published by Bliss (1935), in which the numbers of beetles killed after 5 hour exposure to carbon disulphide at N = 8 different concentrations are recorded. Denoting the number of beetles killed at, and exposed to, dose $x_i, i = 1, \dots ,8$, by y_i and n_i respectively, we fit the following logistic model:

$$y_i \sim Binom(p_i, n_i)$$

$$logit(p_i) = \beta_0 + \beta_1(x_i - \bar{x})$$

Using vague priors on the $\beta's$ of the form $\sim N(0, \sigma = 100^2)$ and ***centering the independent variable***, the following script with missing code chunks is suggested to analyze the data.

```
model {
for (i in 1:8) {
y[i] ~ dbin(p[i], n[i])
…

phat[i] <- y[i]/n[i]
yhat[i] <- n[i]*p[i]
}
…
beta1 ~ dnorm(0, 0.0001)
}
```

```
Data:
list(x = c(1.6907, 1.7242, 1.7552, 1.7842, 1.8113, 1.8369, 1.8610, 1.8839),
n = c(59, 60, 62, 56, 63, 59, 62, 60),
y = c(6, 13, 18, 28, 52, 53, 61, 60))
```

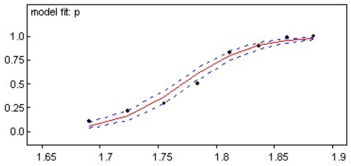
```
Inits:
list(beta0 = …, beta1 = …)
```

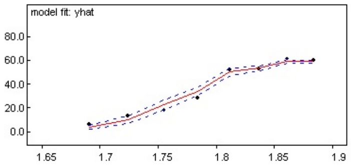
a) Using `ggplot` make appropriate plots of the data.
b) Complete/correct the code above and make a JAGS script to find posterior MCMC for the $\beta's$ and derived nodes p and yhat.
c) Give the MCMC diagnostics for 3 chains and 5000 iterations on each.
d) Make a JAGS script that ***does not center*** the independent variable (Dose)
e) Run the code and show MCMC diagnostics – what do you conclude?
f) Using the `ggmcmc` package make a pairs plot
    a. of the posterior $\beta's$ - centered x
    b. of the posterior $\beta's$ - non-centered x
    c. Compare the pictures and make some conclusions – what does centering accomplish?
    Hint: `ggs_pairs(S, lower = list(continuous = "density"))` see http://xavier-fim.net/packages/ggmcmc/

g) Say which are random:
   a. yhat[i]
   b. phat[i]
   c. beta1
   d. beta0
   e. p[i]
   f. n[i]

h) In the model above you used a logit link – what other links could you use?
   a. With centered data use a different link within your model
   b. Any difference in the conclusions?

i) Duplicate the pictures below in R by making your own script that will take the data and MCMC output (from model with centered x, logit link) and make the plots (these should be far more sophisticated and clear). The plots are p Vs Dose and yhat Vs Dose.





3. Now you will need to analyze the Titanic data set. I want you to perform a logistic regression where "Survived" is the response. Please note that this question is open for you to be creative and answer as best you can. Show me what you can do!!
   1. Describe the data – that is give a full description of the variables. See https://www.youtube.com/watch?v=49fADBfcDD4&t=3401s for help.
   2. Plot the data in at least four useful ways using ggplot. Make sure you describe the plots.
   3. Make a JAGS script to analyze the data using whatever x variables you wish – make different linear predictors and use DIC to choose between the models.
   4. Make conclusions about the probability of survival based on different combinations of independent variables.