# ST440/540 Applied Bayesian Analysis
## Lab activity for 3/18/2024

### Welcome back announcements:

- A5 is due on Friday
- E2 is due April 15 and will be assigned about a week prior. It will have same format as E1 except I will provide a dataset.
- Final project groups are posted on moodle. Please set up a time to meet as soon as possible, if you need help contacting your group members email bjreich@ncsu.edu

## A. HOMEWORK AND CLASS PARTICIPATION SOLUTIONS

Why is the effective sample size of an MCMC chain usually less than the length of the chain? The ESS is smaller because it discounts the samples for being correlated and thus at least partially redundant.

## B. DISCUSSION QUESTIONS

(1) In this problem we will analyze the data from this old exam.

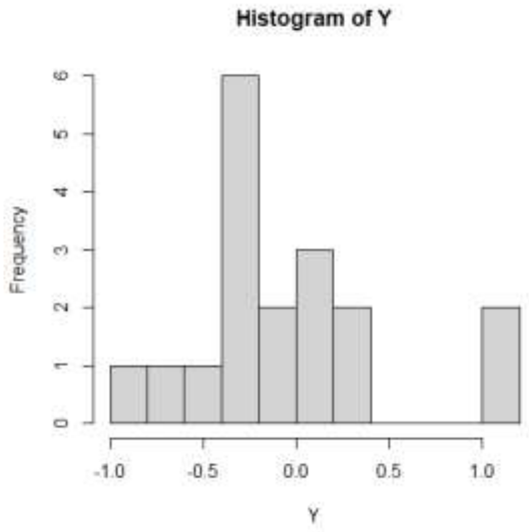
https://st540.wordpress.ncsu.edu/files/2022/02/E1_2022.pdf

In this analysis, each county's data is summarized by the difference in log medal rates between years,

Y = log(Y1/n1) – log(Y0/n0) = log[(Y1/n1)/(Y0/n0)].

The data and a plot are given below.

```
y0 <- c(24,11,25,18,1,26,5,125,94,19,4,108,41,13,63,47,17,41)
n0 <- c(129,81,135,162,94,275,208,410,396,175,229,545,417,140,384,304,236,395)
y1 <- c(22,35,36,29,9,40,11,195,174,33,22,101,58,16,100,65,19,51)
n1 <- c(258,294,280,328,275,423,385,489,522,401,422,647,617,426,599,530,462,621)
lograte0 <- log(y0/n0)
lograte1 <- log(y1/n1)
Y <- lograte1 - lograte0
> round(Y,2)
[1] -0.78 -0.13 -0.36 -0.23 1.12 0.00 0.17 0.27 0.34 -0.28 1.09 -0.24
[13] -0.04 -0.91 0.02 -0.23 -0.56 -0.23
hist(Y,breaks=10)
```



(a) Describe a Bayesian t-test to determine if the mean difference in log rates is greater than zero. Give the likelihood, prior and formula for the posterior, and describe how you would conduct the hypothesis test.

> $Y|\mu,\sigma^2$ ~ Normal($\mu$, $\sigma^2$) with Jefferies prior $\pi(\mu, \sigma_2) = (1/ \sigma^2)^{3/2}$. The posterior is $\mu$ |Y ~ $t_n(\bar{Y},s^2/n)$ where $\bar{Y}$ and $s^2$ are the sample mean and variance of $Y_1,\ldots,Y_n$. We then compute the posterior probability that $\mu > 0$ and conclude there is an advantage if this exceed 0.9.

(b) What are the main assumptions of this analysis? Do you think they are justified? How would you check?

- The data are Gaussian, probably OK for large counts, could check a histogram (it's below)
- Mean is constant over time, we will check below
- Independence over time, could look at an ACF

(2) Next we will analyze the Olympics data to test whether home-county advantages changes over time. We will conduct a simple linear regression with time as the covariate.

(a) Write the model being fit in the code below in mathematical notation.

The covariate $X_i$ is the year of the $i^{th}$ Olympics. The response model is

$Y_i \sim \text{Normal}(a+bX_i, \sigma^2)$

where the observations are independent across i.

(b) Give an interpretation of the parameters a and b.

- The intercept a is the mean value with X=0 AD, which is not interesting (except maybe in the history department)
- The slope b is the increase in the expected difference in the difference in log rates per year

(c) Evaluate the convergence of the MCMC chains

It's bad! I would not trust these results.

(d) Summarize the results of the analysis, i.e., does the home-county advantage change over time?

Since the 95% interval for b includes zero, we cannot conclude the home-county changes over time. Although, this is questionable because of poor convergence.

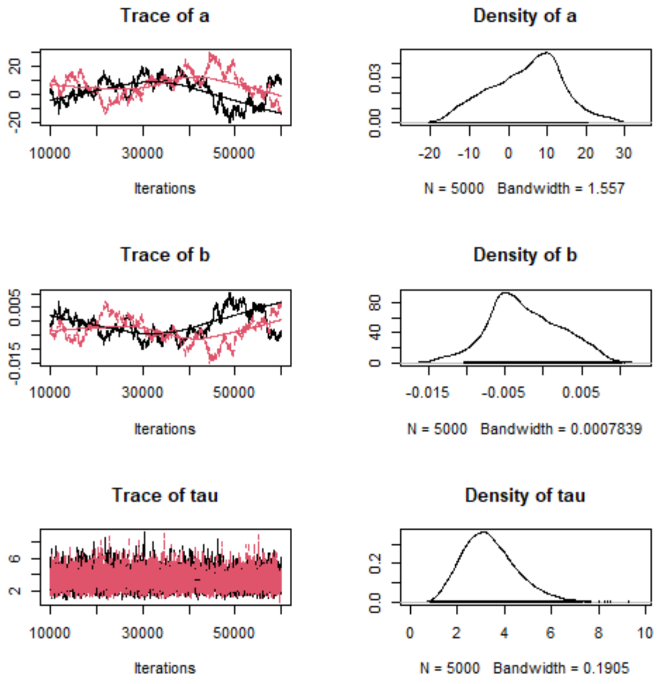
```
X <- seq(1952,2020,4) # treat the year as a covariate
X
[1] 1952 1956 1960 1964 1968 1972 1976 1980 1984 1988 1992 1996 2000 2004 2008 2012 2016 2020

model_string <- textConnection("model{
   for(i in 1:n){
     Y[i]   ~ dnorm(mu[i],tau)
     mu[i] <- a+b*X[i]
   }
   a   ~ dnorm(0,0.001)
   b   ~ dnorm(0,0.001)
   tau ~ dgamma(0.1,0.1)
}")

# Fit the model

inits   <- list(tau=1)
data    <- list(Y=Y,X=X,n=length(Y))
model   <- jags.model(model_string,data = data, inits=inits, n.chains=2, quiet=TRUE)
update(model, 10000, progress.bar="none")
samples <- coda.samples(model, variable.names=c("a","b","tau"),
                        n.iter=50000, progress.bar="none", thin=10)

# Plot the results
plot(samples)
```



(3) The analysis below is identical to the previous analysis except that the covariate is scaled. The code is identical after the definition of the covariate

```
> X <- seq(1952,2020,4) # treat the year as a covariate
> X <- as.vector(scale(X))
> round(X,2)
 [1] -1.59 -1.40 -1.22 -1.03 -0.84 -0.66 -0.47 -0.28 -0.09  0.09  0.28  0.47
[13]  0.66  0.84  1.03  1.22  1.40  1.59
> mean(X)
[1] 0
> sd(X)
[1] 1
```

(a) Given an interpretation of a and b in this model.

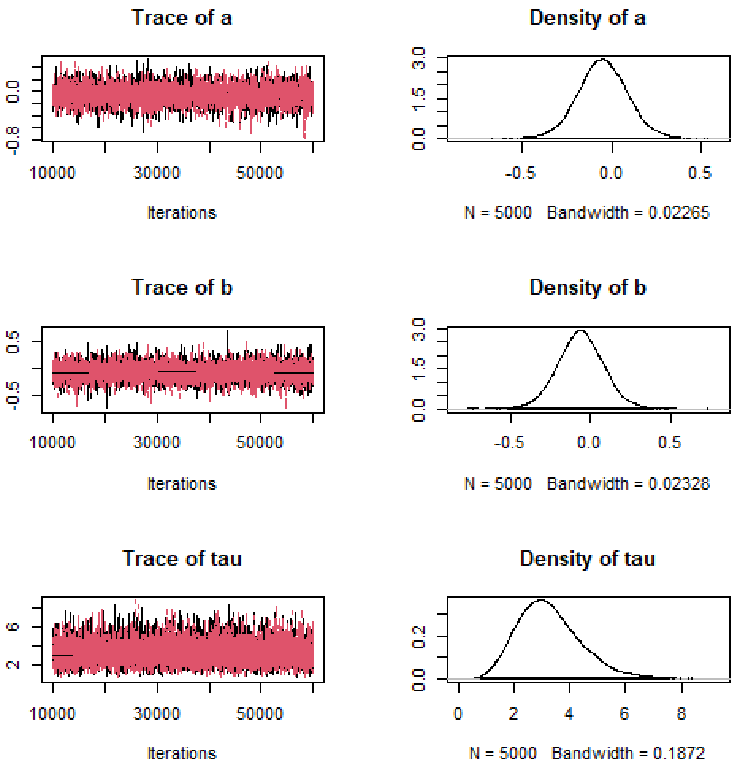
- The intercept a is the mean value with X=0, which is the mean year, E(X)=1986
- The slope b is the increase in the expected difference in the difference for every sd(X)=21 years.

(b) Assess convergence.

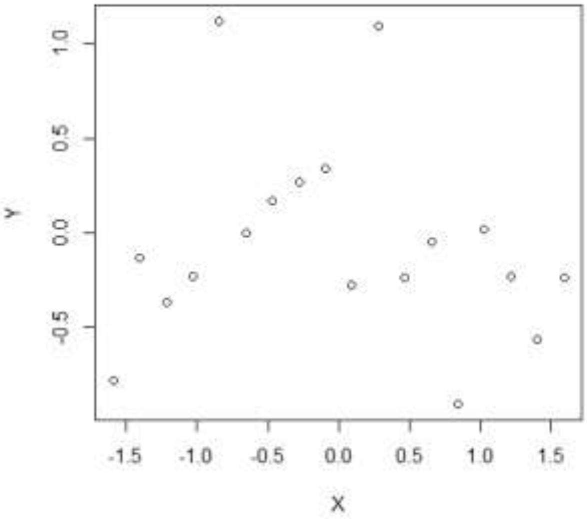
Beautiful!

(c) Does home-county advantage change over time?

No, the 95% interval for b includes zero.



(4) Here is a plot of the data from the previous analysis.



Describe the main assumptions of the simple linear regression model and how you might check that they are appropriate.

The assumptions of linear regressions are
(1) Mean is linear: could plot the residuals ($Y_i$ – posterior mean of a – $X_i$posterior mean of b) versus X and see if there is a pattern.
(2) Variance is the same for all residuals: could plot the sample residual variance for different time periods
(3) Residuals are independent: compute ACF of the errors
(4) Residuals are Gaussian: Make a histogram of the residuals

(5) You are working at a large clinic in Raleigh and you collect data for n patients. The response, $Y_i$, is total amount of money they charged insurance in the past 5 years. The covariates $X_{i1},\dots,X_{ip}$ are features of the patient. Assume the multiple linear regression model:

$$Y_i = b_0 + X_{i1}b_1 + \dots + X_{ip}b_p + e_i$$

(a) Give a scenario where the following priors might be useful:

(i) Jeffries prior
(ii) Bayesian LASSO prior
(iii) An informative prior with $b_j \sim \text{Normal}(m_j,v)$ with small variance v

(i) We have no prior info and p is small compared to n, so maybe we use only a few SES variables as predictors
(ii) We have no prior info and p is large, so maybe we use genetic predictors
(iii) Maybe we have a study from another (we believe to be similar) branch that had a large sample, and the new clinic in Raleigh doesn't have a lot of data. (This is often called transfer learning)

(b) For a given data analysis, how might you select between these priors? That is, what numerical criteria would you use to determine which model is preferred?

This will be the topic of a next section on model selection and diagnostics, but a simple method is cross validation (i.e., fit the model on a subset of data and see how will it predicts the other observations).