ST440/540 Applied Bayesian Analysis
Lab activity for 3/17/2025

Welcome back announcements:

- Q9 is due Friday
- A5 is due on Friday
- E2 is due April 14 and will be assigned about a week prior. It will have same format as E1.
- Final project groups are posted on moodle. Please set up a time to meet as soon as possible. Email me if you have trouble contacting your group.

A. STUDENT QUESTIONS

None, Spring Break

B. HOMEWORK AND CLASS PARTICIPATION SOLUTIONS

None, Spring Break

C. DISCUSSION QUESTIONS

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

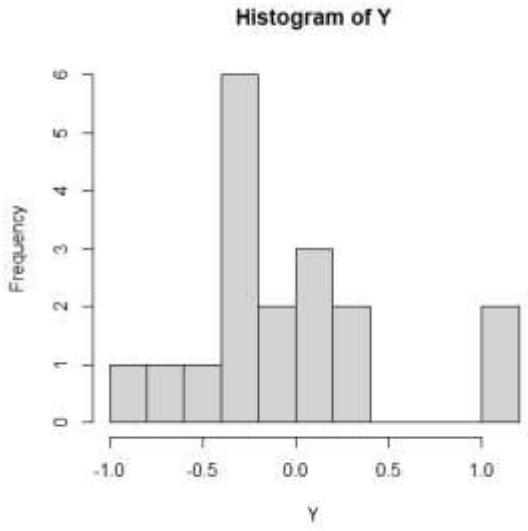
https://st540.wordpress.ncsu.edu/files/2022/02/E1_2022.pdf

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.

$Y|\mu,\sigma^2 \sim$ Normal(μ, σ^2) with Jefferies prior $\pi(\mu, \sigma_2) = (1/\sigma^2)^{3/2}$. The posterior is $\mu|Y \sim t_n(\bar{Y}, s^2/n)$ 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) The code at the end of the document loads the data from

Cloud KA, BJ Reich, CM Rozoff, S Alessandrini, WE Lewis, L Delle Monache, 2019: A Feed Forward Neural Network Based on Model Output Statistics for Short-Term Hurricane Intensity Prediction. Wea. Forecasting, 34, 985–997, https://doi.org/10.1175/WAF-D-18-0173.1.

The response variable (Y) is the observed hurricane intensity (knots) and the covariate (X) is a prediction made 12 hours earlier.

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

The simple linear regression model is

$$Y_i = a + bX_i + e_i$$

where $e_i \sim Normal(0, \sigma^2)$, independent for i=1,…,n.

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

The intercept a is the expected wind speed when the forecast is zero, and the slope b is the increase in the expected wind speed for an increase of one knot in the forecast.

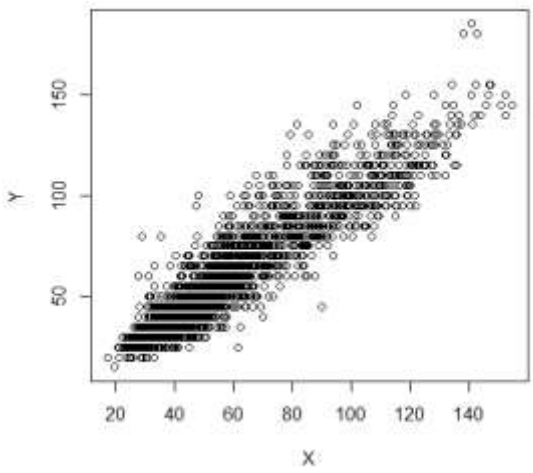
(c) Evaluate the convergence of the MCMC chains

Excellent.

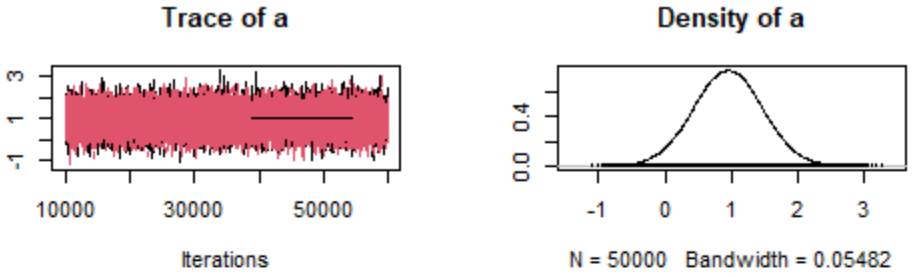
(d) Summarize the results of the analysis, i.e., is the forecast effective?

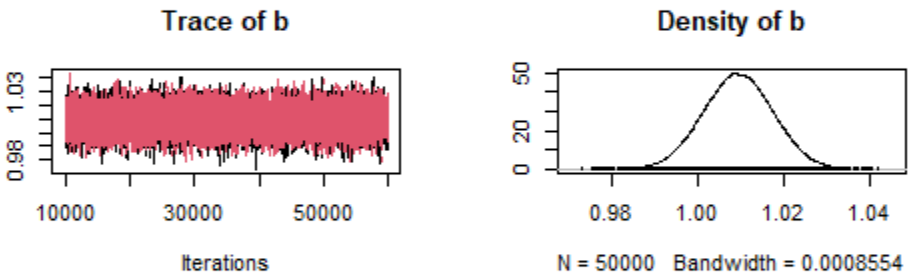
The forecast appears to be effective. The posterior of the slope is far from zero so the predictor is statistically significant. The slope is near one and the intercept is near zero, which when X is a forecast of Y implies the forecast does not have substantial bias.

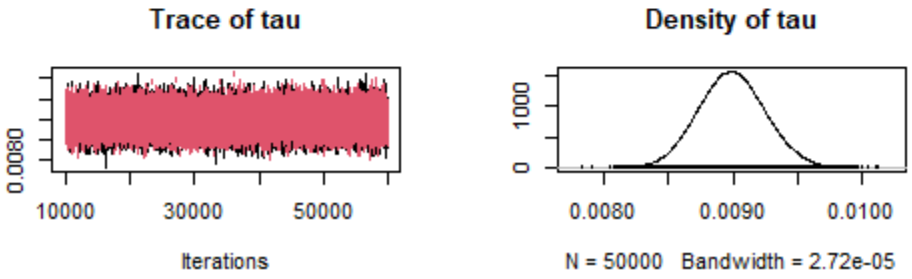
```
X <- HWRF
plot(X,Y)
```



```
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=1)

# Plot the results
plot(samples)
```







(3) List 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 X values

(3) Residuals are independent: hard to test in generally, here you might look for trends within a storm

(4) Residuals are Gaussian: Make a histogram of the residuals

(4) The analysis below adds another forecast to the model. This forecast is the official forecast from the National Hurricane Center (NHC)

(a) Write the model in mathematical notation

The multiple linear regression model is

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + e_i$$

where $e_i \sim Normal(0, \sigma^2)$, independent for i=1,…,n.

(b) Give an interpretation of b1 in this model

The slope b1 is the increase in the expected wind speed for an increase of one knot in the HWRF forecast while holding the NHC constant.

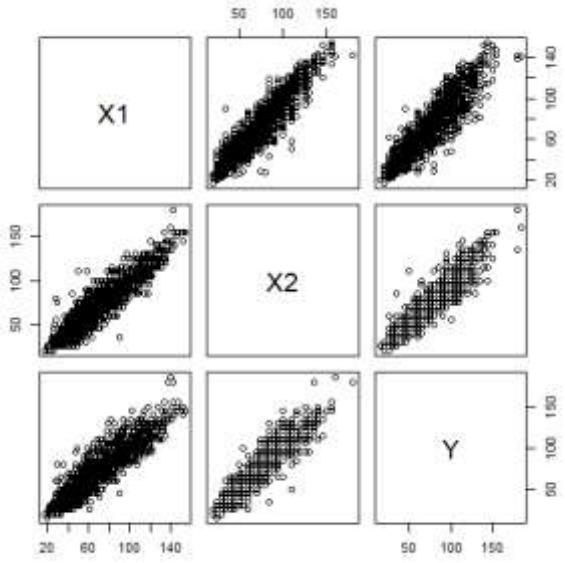
(c) Assess convergence.

Convergence is pretty good, but worst than the simple regression because of correlation between the covariates (collinearity).

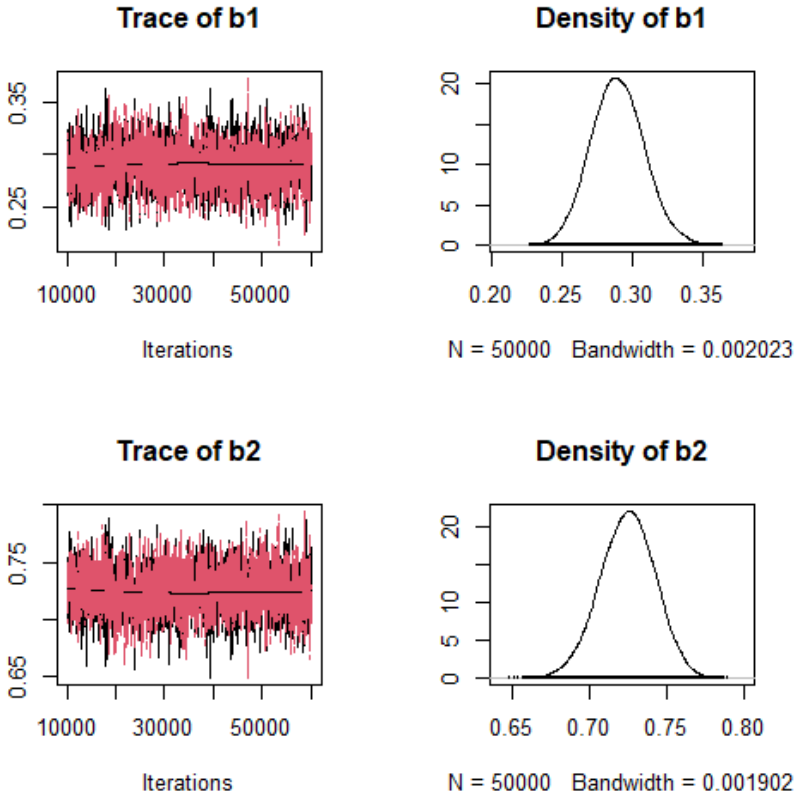
(d) Does adding the second forecast improve the model? How might you show this?

Both covariates have 95% intervals that exclude zero so this is some evidence that the second forecast improves the model, but cross-validation would be more persuasive.

```
X1 <- HWRF
X2 <- NHC
pairs(cbind(X1,X2,Y))
```




```
model_string <- textConnection("model{
   for(i in 1:n){
     Y[i]   ~ dnorm(mu[i],tau)
     mu[i] <- b0 + b1*X1[i] + b2*X2[i]
   }
   b0   ~ dnorm(0,0.001)
   b1   ~ dnorm(0,0.001)
   b2   ~ dnorm(0,0.001)
   tau  ~ dgamma(0.1,0.1)
}")

# Fit the model

inits   <- list(tau=1)
data    <- list(Y=Y,X1=X1,X2=X2,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("b1","b2"),
                        n.iter=50000, progress.bar="none", thin=1)

# Summarize the results
summary(samples)

     2.5%    25%    50%    75%  97.5%
b1 0.2532 0.2767 0.2894 0.3024 0.3286
b2 0.6886 0.7128 0.7252 0.7371 0.7591

plot(samples)
```


Trace of b1
0.35
0.25
10000
30000
50000
Iterations
Density of b1
20
10
5
0
0.20
0.25
0.30
0.35
N = 50000 Bandwidth = 0.002023
Trace of b2
0.75
0.65
10000
30000
50000
Iterations
Density of b2
20
10
5
0
0.65
0.70
0.75
0.80
N = 50000 Bandwidth = 0.001902

(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 $X_{i1},\ldots,X_{ip}$ are features of the patient. Assume the multiple linear regression model:

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

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

(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).

LOADING THE HURRICANE DATA

```
library(rjags)
library(lubridate)

filename   <-
"https://www4.stat.ncsu.edu/~bjreich/BSM2/Chapter8/cloud_data_clean.csv"
dat        <- read.csv(url(filename))
complete   <- rowSums(is.na(dat))==0
dat        <- dat[complete,]
ID         <- dat$StormID
lead_time  <- dat$lead_time
basin      <- ifelse(dat$basin=="atlantic",1,0)
X          <- as.matrix(dat[,5:22])
Xs         <- scale(X)
HWRF       <- dat$HWRF
NHC        <- dat$NHC
Y          <- dat$VMAX
year       <- year(dat$Date)
lead_times <- sort(unique(lead_time))
nleads     <- length(lead_times)

Y          <- Y[lead_time==12]
HWRF       <- HWRF[lead_time==12]
NHC        <- NHC[lead_time==12]
```