ST440/540 Applied Bayesian Analysis
Lab activity for 4/14/2025

Announcements

- No in-person class on Monday. We will have a zoom office hour to discuss the 440 exam.
- All final projects will be a zoom recording, we will not meet in person. I will send more info about this soon.
- I will email an exam solution soon.

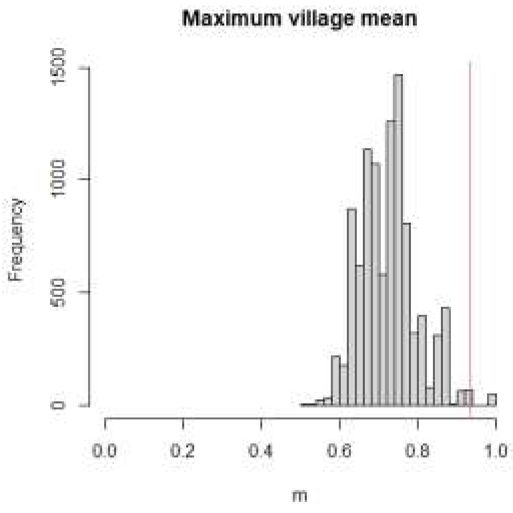
A. HOMEWORK AND QUIZ SOLUTIONS

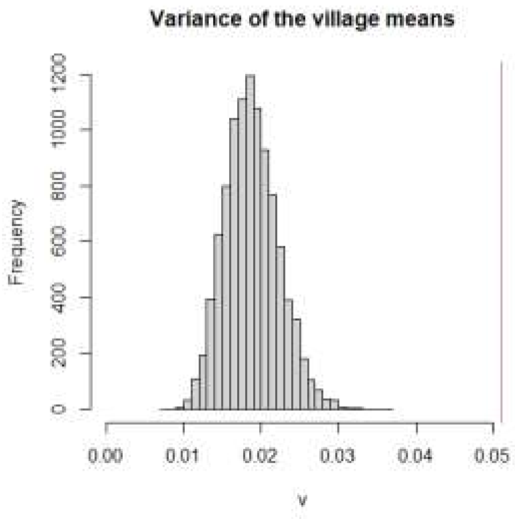
Chapter 5, problem 6

There are of course many other (simpler) options, but since we previously included village random effects and here we are dropping those, I chose to test this assumption. For each village I computed the sample proportion, and then used the maximum and variance of these sample proportions as the criteria. It turns out the model without random effects does not fit well based on these criteria, so we should probably add the village random effects back to the model.

```
# Load the data
library(geoR)
data(gambia)
Y <- gambia[,3]
X <- scale(gambia[,4:8])
s <- gambia[,1:2]
n <- length(Y)

S <- unique(s) # Lat/long of the villages
m <- nrow(S)
village <- rep(0,n)
for(j in 1:m){
   d             <- (s[,1]-S[j,1])^2 + (s[,2]-S[j,2])^2
   village[d==0] <- j
}

# Fit the model in JAGS
mod <- textConnection("model{
 for(i in 1:n){
 Y[i] ~ dbern(pi[i])
 logit(pi[i]) <- beta[1] + X[i,1]*beta[2]+
                 X[i,2]*beta[3] + X[i,3]*beta[4] +
                 X[i,4]*beta[5] + X[i,5]*beta[6]
 }
 for(j in 1:6){beta[j] ~ dnorm(0,0.01)}
}")
```

```
data  <- list(Y=Y,X=X,n=n)
model <- jags.model(mod,data = data, n.chains=1,quiet=TRUE)
update(model, 5000, progress.bar="none")
beta  <- coda.samples(model, variable.names=c("beta"),
                             n.iter=10000, progress.bar="none")[[1]]


# Village sample proportions
ybar0 <- aggregate(Y~village, FUN=mean)[,2]
m0    <- max(ybar0)
v0    <- var(ybar0)


# Posterior predictive checks
S      <- nrow(beta)
m      <- rep(0,S)
v      <- rep(0,S)

for(i in 1:S){
  b     <- beta[i,]
  eta   <- b[1] + X%*%b[2:6]
  prob  <- exp(eta)/(1+exp(eta))
  y     <- rbinom(n,1,prob)
  ybar  <- aggregate(y~village, FUN=mean)[,2]
  m[i]    <- max(ybar)
  v[i]    <- var(ybar)
}

hist(m,breaks=25,main="Maximum village mean",xlim=0:1)
abline(v=m0,col=2)

hist(v,breaks=25,main="Variance of the village means",xlim=c(0,v0))
abline(v=v0,col=2)
```





Chapter 5, problem 8

Again, there are many approaches here. This is a small problem with only 10 observations, so I decided to use a discrepancy measure for each observation. The measure is whether the draw from the prediction distribution is greater than the actual observation. None of the means of these measures are close to zero or one, so by this measure the model seems to fit OK.

```
Library(rjags)

# Load the data
Y <- c(64, 72, 55, 27, 75, 24, 28, 66, 40, 13)
N <- c(75, 95, 63, 39, 83, 26, 41, 82, 54, 16)
q <- c(0.845, 0.847, 0.880, 0.674, 0.909, 0.899, 0.770, 0.801, 0.802, 0.875)
X <- log(q)-log(1-q) # X = logit(q)
inits <- c("RW","JH","KL","LJ","SC","IT","GA","JW","AD","KD")

# Fit the model in JAGS
model_string <- textConnection("model{
   # Likelihood
    for(i in 1:10){
      Y[i]        ~ dbinom(p[i],N[i])
      logit(p[i]) <- beta[1] + beta[2]*X[i]
    }
   # Priors
    beta[1] ~ dnorm(0,0.01)
    beta[2] ~ dnorm(0,0.01)

   # PPD
   for(i in 1:10){
      Yp[i]    ~ dbinom(p[i],N[i])
      D[i]     <- step(Yp[i]-Y[i])
   }
}")

data   <- list(Y=Y,N=N,X=X)
model <- jags.model(model_string,data = data, n.chains=2,quiet=TRUE)
update(model, 10000, progress.bar="none")
samps <- coda.samples(model, variable.names=c("D"), thin=5,
                      n.iter=20000, progress.bar="none")

# Summarize the PPD
stats <- summary(samps)$stat
rownames(stats)<-inits
round(stats,3)

    Mean    SD Naive SE Time-series SE
RW 0.301 0.459    0.005          0.005
JH 0.942 0.234    0.003          0.003
KL 0.456 0.498    0.006          0.006
LJ 0.344 0.475    0.005          0.008
SC 0.446 0.497    0.006          0.007
IT 0.390 0.488    0.005          0.006
GA 0.828 0.377    0.004          0.005
JW 0.315 0.465    0.005          0.005
AD 0.755 0.430    0.005          0.005
KD 0.796 0.403    0.005          0.005
```

B. DISCUSSION QUESTIONS

(1) Discuss your exam solution with your group. What did you do? Did it work?

(2) The following two questions are about missing data.

(a) Say you're are doing a Bayesian simple regression of Y onto X and observations are missing X, Y, neither or both. Which observations can you safely discard and why?

Only observations with both X and Y are informative for their correlation, so the others can be discarded.

(b) Say you're are doing a Bayesian multiple regression of Y onto X1 and X2 and observations are missing X2, Y, neither or both. Which observations can you safely discard and why?

All observations except those with only X1 have some information. If Y missing you can still use the observation to model X2|X1 for imputation, and if X2 is missing you can learn about the correlation between X1 and Y.

(3) Causal inference can be viewed as a missing data problem. Say your company has n employees. For a given employee, let X be the number of years with the company prior to 2022 and A=1 if they choose to participate in an online training session and A=0 otherwise. For each employee we envision two potential outcomes

- Y(0) is the 2022 performance score if they do not take the training
- Y(1) is the 2022 performance score if they take the training

Y is the observed value of the 2022 score, so either Y(0) or Y(1) depending on A, i.e., Y=Y(A). The causal effect is the average of Y(1)-Y(0) over the n employees

Here are a few fake observations. Plots are on the next page

Employee	X	A	Y(0)	Y(1)
1	3.2	0	73	NA
2	7.1	0	61	NA
3	1.0	1	NA	95

(a) Explain why a Bayesian t-test comparing the average of Y for the employees that did and did not take the training in a biased estimate of the causal effect.

It ignores lurking variable X. X is a lurking variable because the plots show it is correlated with by A and Y.

(b) Looking at the scatterplot below, would you say that the training session is beneficial?

It appears to improve performance for the new employees (small X).

(c) How might you fill in the missing observations in the table above?

Run a multiple regression with Y as the response and A and X as the predictor, and then make predictions with the non-observed A (i.e., 1-A) as the covariate. Or fit a regression separate for A=0 and A=1.

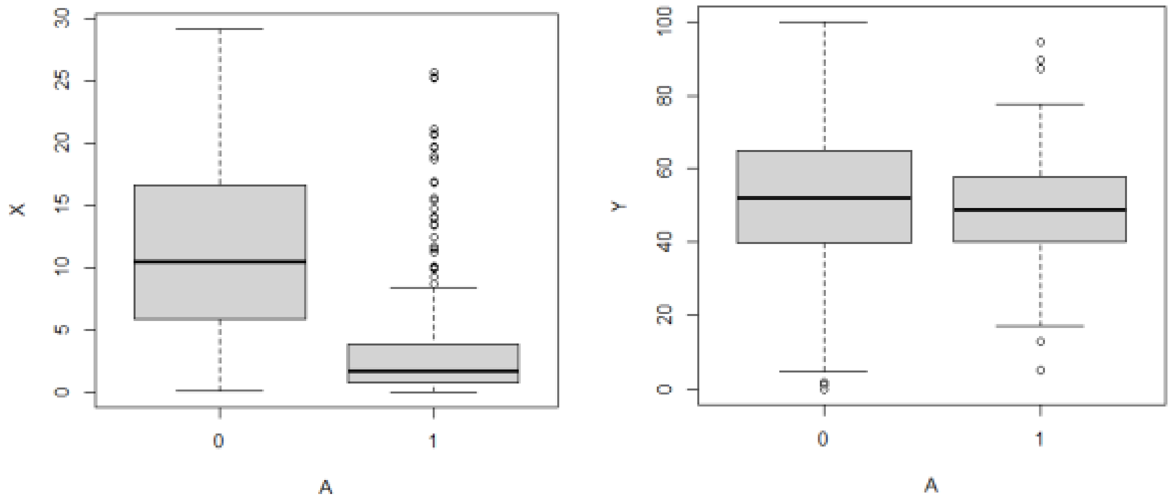
(d) Based on this completed table, how would you estimate the causal effect and test if it is positive?

(1) For each MCMC iteration fill in the missing values as draws from the PPD
(2) For each iteration, compute the mean of the differences in the last two columns
(3) Compute a 95% interval for the causal effect.

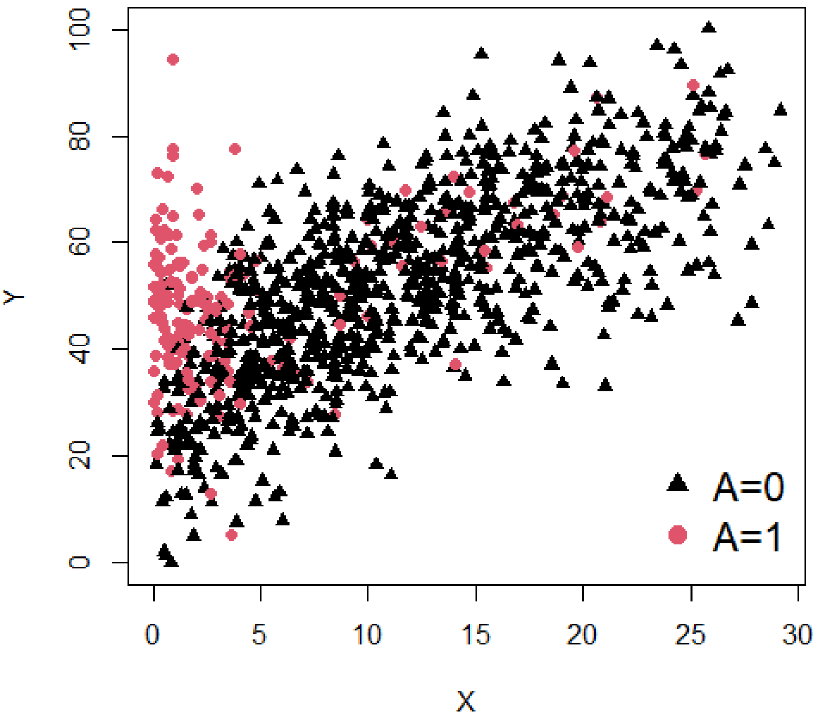
(e) What are the main assumptions you're making and how might you verify them?

There are no unobserved lurking variables. Also, we are assuming the stat model for imputation is correct (normal, linear, whatever else we assume).

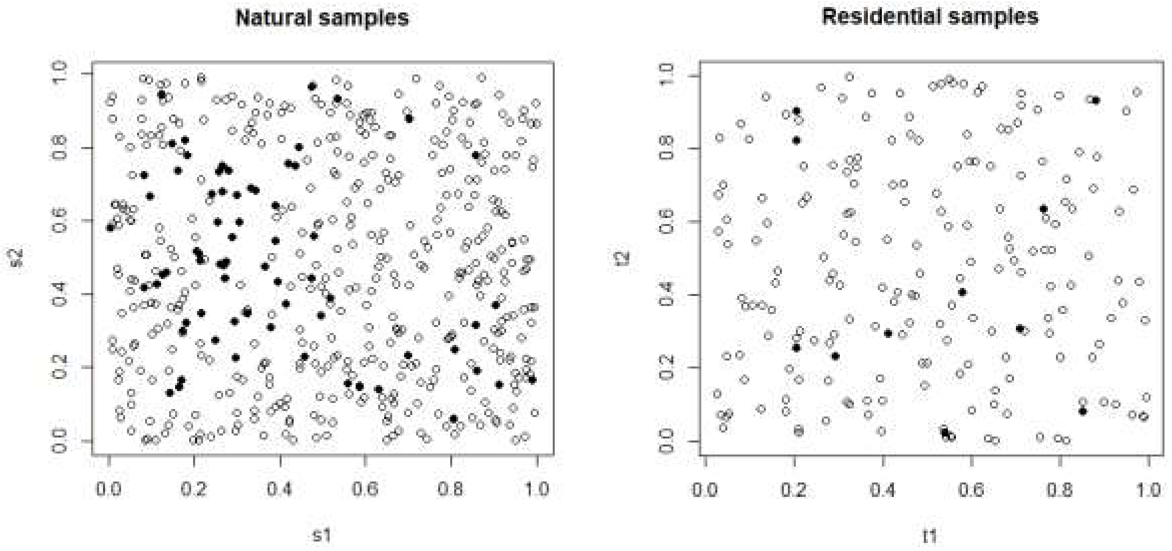
Plot of X and Y for the those that did (A=1) and did not (A=0) take the training.



Plot of the performace scores Y by X and A.



(4) Nontuberculous mycobacteria (NTM) occurs in the environment and is harmful to humans. The mechanism by which it moves from soil and water to humans remains uncertain. The data you are analyzing comes from two sources. The first is soil samples. For sample i, let Y_i =1 if NTM is found and Y_i=0 otherwise, s_i be the spatial location (lat/long) of sample and X_i be environmental covariates such as elevation, soil type, etc. The second data source is human samples. Let Z_k = 1 if they have NTM and Z_k = 0 otherwise and t_k be the spatial location of their residence. Fake data is below.



(a) The two data sources are collected at different locations. Specify a model to predict the presence of natural NTM (s) at the locations of the residential samples (t).

Logit(P(Y_i=1)) = beta0 + lat_i*beta1+long_i*beta2+elev_i*beta3…

(b) Given the status (Y=0 or 1) the natural sample at the residential sample, specify a model for the residential data that would allow you to test whether NTM in the local environment is predictive of human disease. What are the main assumptions you're making? Any other data you would like to collect?

Logit(P(Z_i=1)) = alpha0 + lat_i*alpha1+long_i*alpha2+Y_i*alpha3…

(c) Describe how you would fit a hierarchical model for both data sources simultaneously that would properly account for uncertainty in your imputation of NTM from s to t.

Fit a model in JAGS that simultaneously model Y and Z, with the missing Y's at t entered as NA.

(d) How would a frequentist do (c)?

(i) Plug in, i.e., take fitted values of Y as known covariates for Z.
(ii) Do (i) a bunch of times. This is called multiple imputation.