14.1

** Problem 1: 14.1 BDA ** Setting up the Observations Yi and Explanatory Variables Xij:

```
library(MASS)
y_1 <- c(5.0, 13.0, 7.2, 6.8, 12.8, 5.8, 9.5, 6.0, 3.8, 14.3, 1.8, 6.9, 4.7, 9.5)
y_2 <- c(0.9, 12.9, 2.6, 3.5, 26.6, 1.5, 13.0, 8.8, 19.5, 2.5, 9.0, 13.1, 3.6, 6.9)
y_3 <- c(14.3, 6.9, 7.6, 9.8, 2.6, 43.5, 4.9, 3.5, 4.8, 5.6, 3.5, 3.9, 6.7)
y <- c(y_1, y_2, y_3) # Radon Measurements

# Counties
County_1 <- c(rep(c(1), times = length(y_1)),rep(c(0), times = length(y_2)),rep(c(0),tim
es = length(y_3)))
County_2 <- c(rep(c(0), times = length(y_1)),rep(c(1), times = length(y_2)),rep(c(0),tim
es = length(y_3)))
County_3 <- c(rep(c(0), times = length(y_1)),rep(c(0), times = length(y_2)),rep(c(1),tim
es = length(y_3)))

# 1: Measurement recorded on Floor 1, 0: Otherwise
Floor1_y1 <- c(1,1,1,1,1,0,1,1,1,0,1,1,1,1)
Floor1_y2 <- c(0,1,1,0,1,1,1,1,1,0,1,1,1,0)
Floor1_y3 <- c(1,0,1,0,1,1,1,1,1,1,1,1,1)
Floor1 <- c(Floor1_y1,Floor1_y2,Floor1_y3)

# Explanatory Variables
X <- cbind(Floor1,County_1,County_2,County_3)
```

- a. Fit a linear regression to the logarithms of the radon measurements in Table 7.3, with indicator variables for the three counties and for whether a measurement was recorded on the first floor. Summarize your posterior inferences in nontechnical terms.

```
n <- nrow(X)
k <- ncol(X)
CondPosterior_sigma <- function(beta){
  s2 <- t((log(y)-(X%*%beta))) %*% (log(y)-(X%*%beta))/(n-k)
  post <- sqrt(((n-k) * s2)/rchisq(1,n-k))
}
CondPosterior_beta <- function(sigma){
  R <- qr.R(qr(X))
  Q <- qr.Q(qr(X))
  beta_hat <- solve(R, t(Q)%*%log(y))
  Vbeta_hat <- solve(R) %*% t(solve(R))
  post_beta <- mvrnorm(1,mu = beta_hat,Sigma = (sigma*Vbeta_hat))
}

nsim <- 1000
sigma <- rep(0, nsim)
beta <- array(0, c(nsim, k))
sigma[1] <- 1
beta[1, ]<- rep(1,k)

for (i in 2:nsim){
sigma[i] <- CondPosterior_sigma(beta[i,])
beta[i,] <- CondPosterior_beta(sigma[i])
}
output <- exp(cbind(beta[ ,2], beta[ ,1]+beta[ ,2], beta[ ,3],
beta[ ,1] + beta[ ,3], beta[ ,4], beta[ ,1] + beta[ ,4], beta[ ,1], log(sigma)))
for (i in 1:ncol(output)) print (round(quantile(output[,i],c(.25,.5,.75)),1))
```

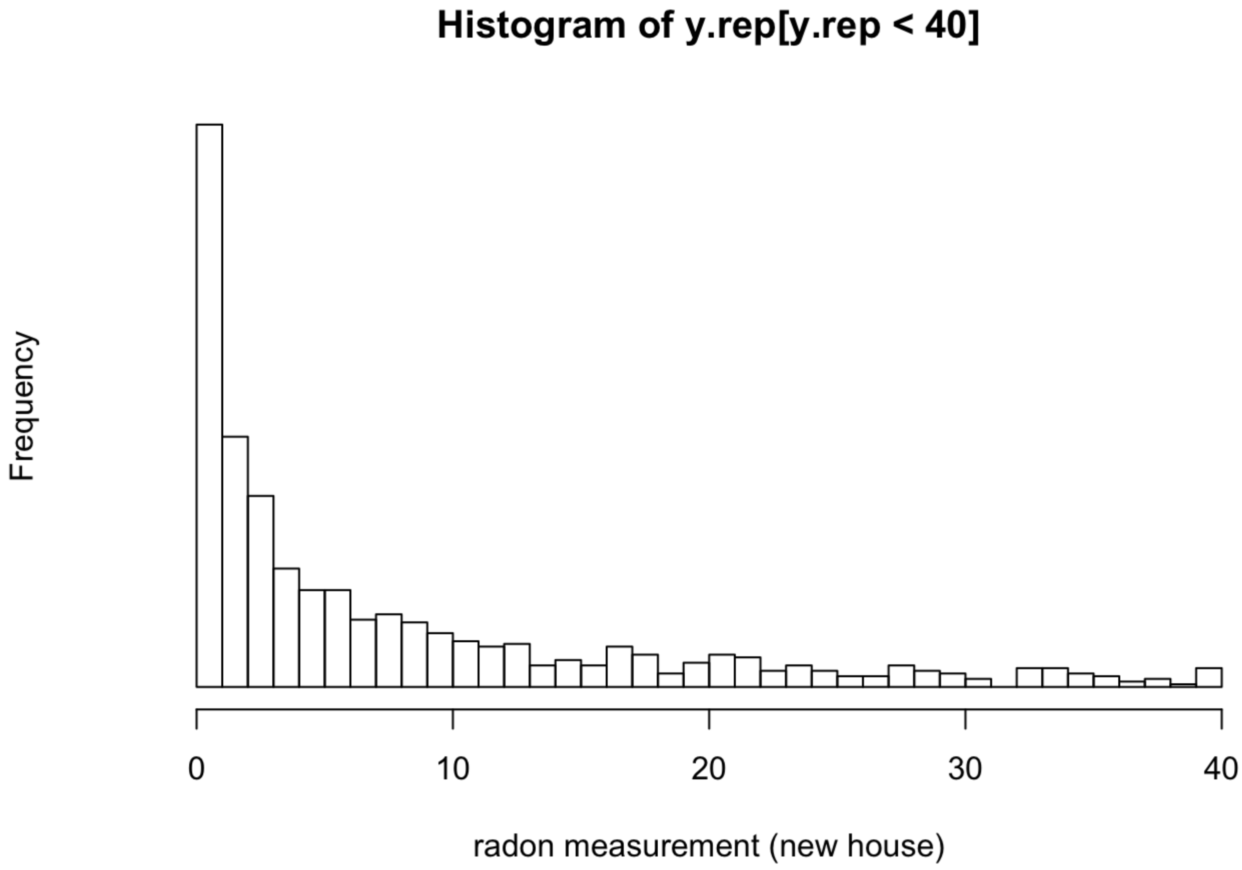
```
## 25% 50% 75%
## 3.2 5.1 8.1
## 25% 50% 75%
## 5.5 6.9 9.1
## 25% 50% 75%
## 3.0 4.6 7.0
## 25% 50% 75%
## 4.8 6.4 8.5
## 25% 50% 75%
## 3.1 4.9 7.6
## 25% 50% 75%
## 5.1 6.8 8.8
## 25% 50% 75%
## 0.9 1.4 2.1
## 25% 50% 75%
## 2.0 2.1 2.3
```

- b. Suppose another house is sampled at random from Blue Earth County. Sketch the posterior predictive distribution for its radon measurement and give a 95% predictive interval. Express the interval on the original (unlogged) scale. (Hint: you must consider the separate possibilities of basement or first-floor measurement.)

```
theta <- rbeta (nsim, 3, 13)
b <- rbinom(nsim, 1, theta)
logy.rep <- rnorm(nsim, beta[,2] + b*beta[,1], sigma)
y.rep <- exp(logy.rep)
print (round(quantile(y.rep,c(.025,.25,.5,.75,.975)),1))
```

```
##  2.5%   25%   50%   75% 97.5% 
##   0.1   1.5   6.6  26.5 535.8
```

```
hist(y.rep[y.rep<40], yaxt="n", breaks=0:40,xlab="radon measurement (new house)", cex=2)
```



- c)

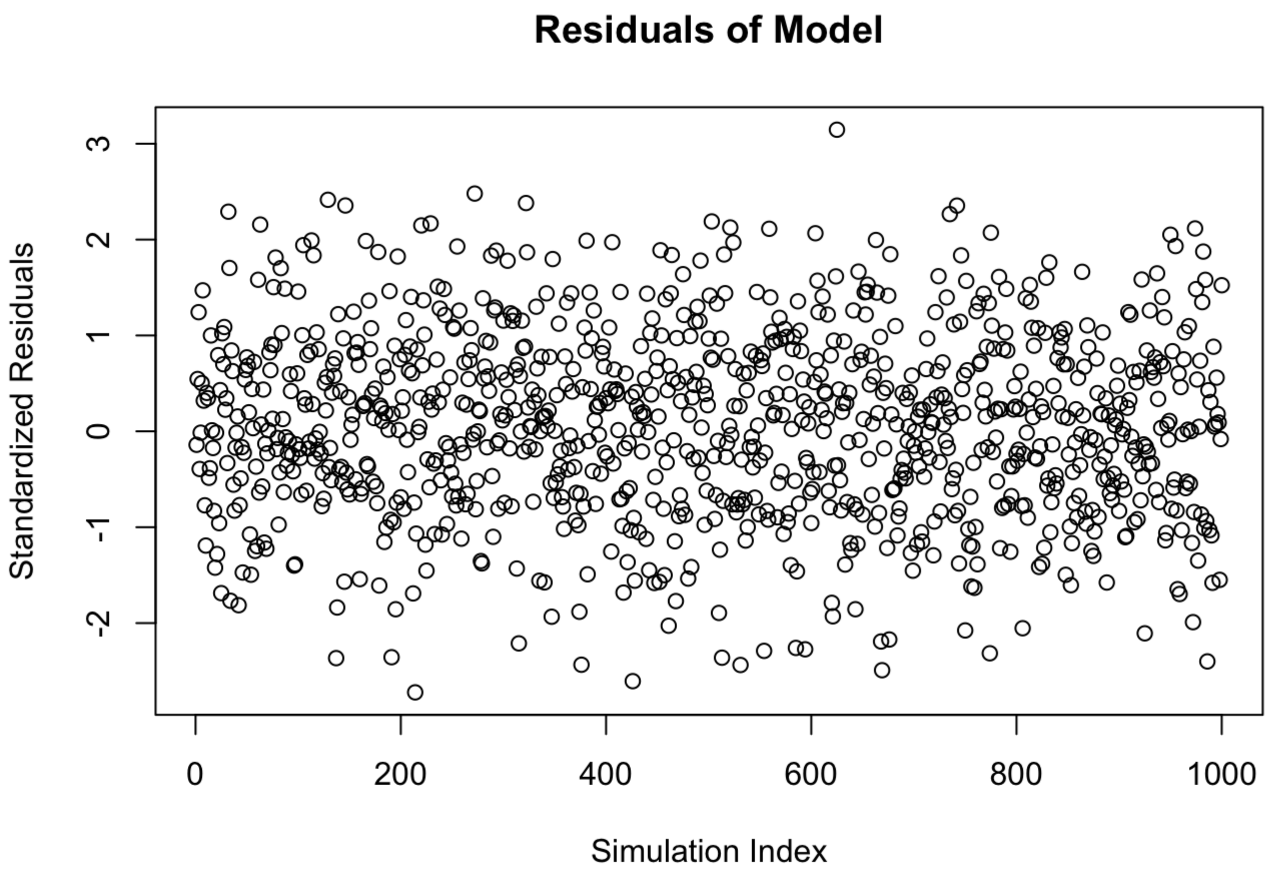
Conduct some posterior predictive checks to assess your model.

```
# overall p-value:
# T(y, t) := |y^(.9) - t^(.5)| - |y^(.1) - t^(.5)|, t^(a) is a-quantile of t
tm <- beta[,2] + b*beta[,1] # t^(.5)
s <- quantile(log(y_1), c(.1, .9))
Tdata <- abs(s[2] - tm) - abs(s[1] - tm)

s <- quantile(logy.rep, c(.1, .9))
Trep <- abs(s[2] - tm) - abs(s[1] - tm)
mean(Trep >= Tdata)
```

```
## [1] 0.235
```

```
standardized_residuals <- (logy.rep - (beta[,2] + b*beta[,1]))/sigma
plot(standardized_residuals,main ='Residuals of Model',xlab='Simulation Index',ylab='Sta
ndardized Residuals')
```



```
outlier_threshold = 3 * sd(log(y_1))
print(sum(standardized_residuals > outlier_threshold)/length(standardized_residuals))
```

```
## [1] 0.048
```