5224-Take-Home-Exam

Sylvia Ye (zy2302)

2017/12/12

Fit a Bayesian linear regression to the logarithms of the radon measurements in Table 7.3 (page 195), with indicator variables for the three counties and for whether a measurement was recorded on the first floor.

1. Report the posterior median and a 50% posterior interval for the mean log radon concentration at each of the six house types (3 counties; measurement taken in basement or first floor). Prepare a summary table and comment on your results.

```
library(mnormt)

# Data for three counties
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)

# Indicator variable for whether a measurement was recorded on the first floor.
# 1 for those on the basement and 0 for those on the first floor.
basement.1 <- c(1,1,1,1,1,0,1,1,1,0,1,1,1,1)
basement.2 <- c(0,1,1,0,1,1,1,1,1,0,1,1,1,0)
basement.3 <- c(1,0,1,0,1,1,1,1,1,1,1,1,1,1)

# Indicator variable for the three counties
counties <- rep(c("Blue Earth", "Clay", "Goodhue"),
                c(length(y.1),length(y.2),length(y.3)))

# Generate the design matrix X and values y
make.indicators <- function (x){
  unique_x <- unique(x)
  mat1 <- matrix (x, nrow=length(x), ncol=length(unique_x))
  mat2 <- matrix (unique_x, nrow=length(x), ncol=length(unique_x), byrow=TRUE)
  (mat1==mat2)*1}
X <- cbind (c(basement.1,basement.2,basement.3), make.indicators(counties))
y <- c(y.1,y.2,y.3)

# Regression_Building Function
Bayes.regression <- function(y, X, Sim.size){
  n <- dim(X)[1]; k <- dim(X)[2];
  m <- lm(y ~ 0 + X)
  beta.hat <- coef(m)
  s <- sigma(m)
  V.beta <- vcov(m) / s^2
  sigma.sim <- s * sqrt((n-k) / rchisq(Sim.size, df=n-k))
  beta.sim <- rep(beta.hat, each=Sim.size) +
                  sigma.sim *rmnorm(Sim.size, mean=rep(0,k),
                  varcov=V.beta)
  answer <- list(beta.sim=beta.sim, sigma.sim=sigma.sim)
  return(answer)
```

```
}

# Do simulation
nsim <- 10000
sim_output <- Bayes.regression(log(y), X, nsim)
beta <- sim_output$beta.sim
sigma <- sim_output$sigma.sim

# the mean log radon concentration for each of the six house types
sim_mean <- cbind (beta[,2], beta[,1] + beta[,2], beta[,3],
                   beta[,1] + beta[,3], beta[,4], beta[,1] + beta[,4])
colnames(sim_mean) <- c("Blue Earth*", "Blue Earth", "Clay*",
                        "Clay", "Goodhue*", "Goodhue")

apply(sim_mean, 2, function(x){round(quantile(x,c(.25,.5,.75)),2)})
```

```
##     Blue Earth* Blue Earth Clay* Clay Goodhue* Goodhue
## 25%        1.39       1.81  1.34 1.72     1.36    1.77
## 50%        1.63       1.95  1.55 1.88     1.59    1.92
## 75%        1.86       2.10  1.76 2.03     1.82    2.07
```

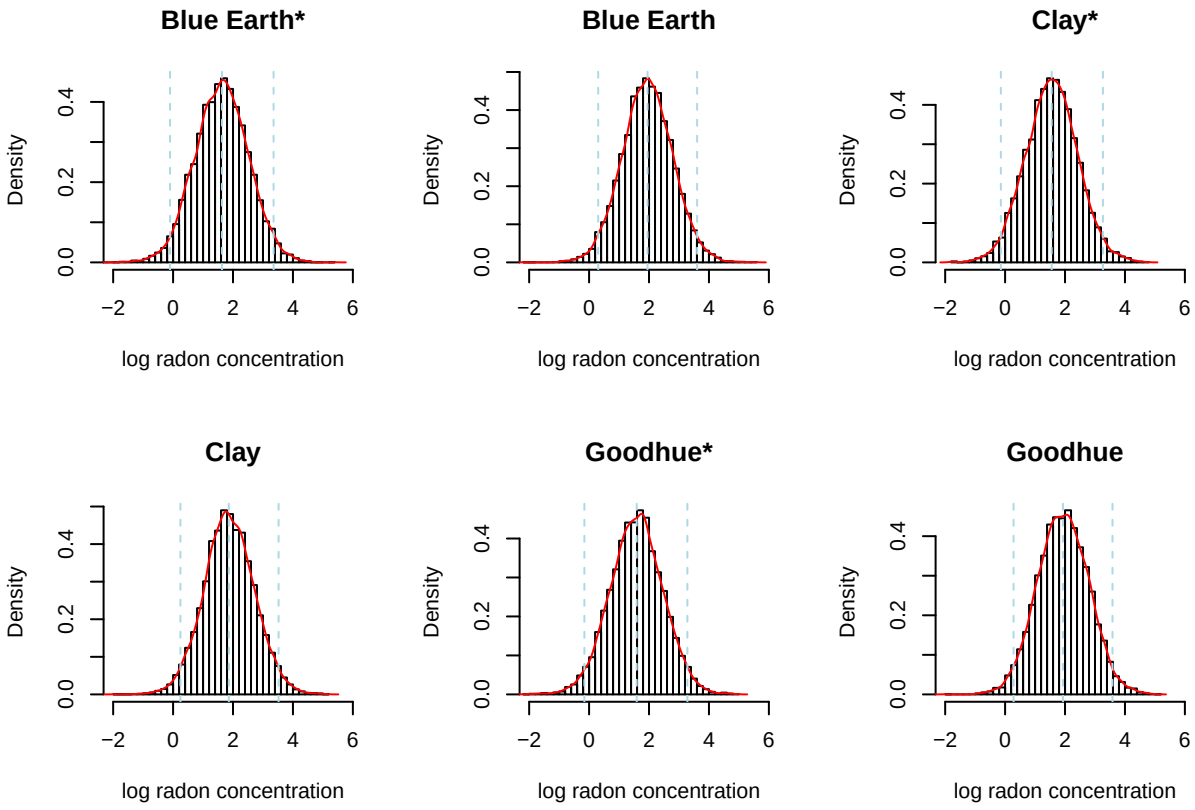
The table above displays the posterior median (50% quantiles) and a 50% posterior interval(from 25% quantiles to 75% quantiles) for the mean log radon concentration at each of the six house types.

Those indicated with asterisks are for the houses on the first floor, and those without asterisks are for the houses on the basement level.

We can see from the table that: 1) Houses on the basement level tend to have higher radon concentration. 2) There isn't very significant difference between these three counties. Relatively, Houses in Clay have lowest radon concentration and houses in Blue Earth have higher radon concentration.

2. Sketch the posterior predictive density for the log radon concentration in a single house, for each of the six house types. Put all six graphs in a single display, using a consistent scale for the x-axes. Comment on your results.

```
logy.rep <- apply(sim_mean, 2, rnorm, n = 10000, sd = sigma)
par(mfrow=c(2,3))
for(i in 1:6){
  hist(logy.rep[,i], freq = FALSE, xlim = c(-2, 6), breaks = 30,
       main = colnames(sim_mean)[i], xlab = "log radon concentration")
  lines(density(logy.rep[,i]), col = "red")
  abline(v = quantile(logy.rep[,i], c(.025, .5, .975)), lty = 2, col = "lightblue")
}
```



The posterior predictive density for the log radon concentration in six house types are similar. Most of the predictive value ranged roughly from 0 to 4 (have larger range than posterior mean).

As for the difference between these types, we can see from the plots that: 1) Houses on the basement level tend to have higher radon concentration. 2) Houses on the basement level's radon concentration tend to have larger range and variance. 3) There isn't very significant difference between these three counties. Relatively, Houses in Clay have lowest radon concentration and houses in Blue Earth have higher radon concentration.

3. Give a 95% predictive interval for the radon concentration in a single house, on the original (unlogged) scale, for each of the six house types.

```
apply(exp(logy.rep), 2, function(x){round(quantile(x, c(.025, .975)),2)})
```

```
##       Blue Earth* Blue Earth Clay*  Clay Goodhue* Goodhue
## 2.5%         0.90       1.35  0.87  1.28     0.85    1.33
## 97.5%       28.63      36.84 26.16 33.82    26.58   36.01
```

The table above displays the 95% posterior interval(from 2.5% quantiles to 97.5% quantiles) for the predictive radon concentration in a single house on the original (unlogged) scale, for each of the six house types.

Those indicated with asterisks are for the houses on the first floor, and those without asterisks are for the houses on the basement level.