Lab 8

Yaqi Shi, 1003813180

2023-03-8

Radon

The goal of this lab is to fit this model to the radon data:

$$y_i|\alpha_{j[i]} \sim N\left(\alpha_{j[i]} + \beta x_i, \sigma_y^2\right), \text{ for } i = 1, 2, \ldots, n$$
$$\alpha_j \sim N\left(\gamma_0 + \gamma_1 u_j, \sigma_\alpha^2\right), \text{ for } j = 1, 2, \ldots, J$$

i.e. varying intercepts, fixed slope on floor. I want you to

- reproduce the graph on slide 47
- plot samples from the posterior predictive distribution for a new household in county 2 with basement level measurement, compared to samples from the posterior distribution of the mean county effect in county 2 (i.e., a graph similar to slide 39).

Here's code to get the data into a useful format:

```
library(tidyverse)
# house level data
d <- read.table(url("http://www.stat.columbia.edu/~gelman/arm/examples/radon/srrs2.dat"), header=T, sep=

# deal with zeros, select what we want, makke a fips variable to match on
d <- d %>%
  mutate(activity = ifelse(activity==0, 0.1, activity)) %>%
  mutate(fips = stfips * 1000 + cntyfips) %>%
  dplyr::select(fips, state, county, floor, activity)

# county level data
cty <- read.table(url("http://www.stat.columbia.edu/~gelman/arm/examples/radon/cty.dat"), header = T, se
cty <- cty %>% mutate(fips = 1000 * stfips + ctfips) %>% dplyr::select(fips, Uppm)

# filter to just be minnesota, join them and then select the variables of interest.
dmn <- d %>%
  filter(state=="MN") %>%
  dplyr::select(fips, county, floor, activity) %>%
  left_join(cty)
head(dmn)
```

```
##    fips              county floor activity     Uppm
## 1 27001 AITKIN                  1      2.2 0.502054
## 2 27001 AITKIN                  0      2.2 0.502054
## 3 27001 AITKIN                  0      2.9 0.502054
## 4 27001 AITKIN                  0      1.0 0.502054
## 5 27003 ANOKA                   0      3.1 0.428565
```

```
## 6 27003 ANOKA                    0      2.5 0.428565
```

Note, in the model:

- y_i is log(activity)
- x_i is floor
- u_i is log(Uppm)

Suggested steps

1. write Stan model (note, you will need samples from post pred distribution, either do in Stan or later in R)

Please see the stan file in the folder

2. Get data in stan format

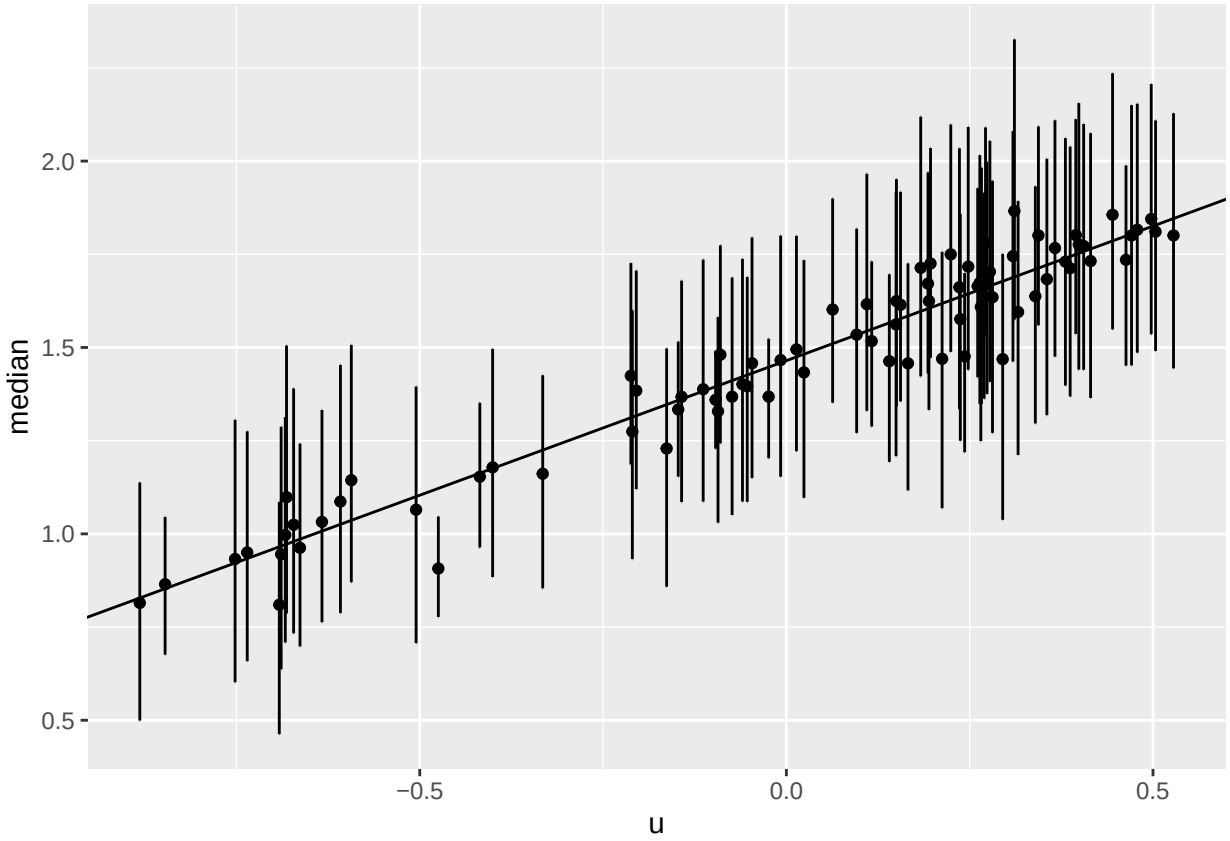
```
library(rstan)

# Set up the data entry
y <- log(dmn$activity)
x <- dmn$floor
u <- log(dmn %>% group_by(county) %>% slice(1) %>% select(Uppm) %>% pull())
N <- nrow(dmn)
J <- length(unique(dmn$county))
county <- as.numeric(as.factor(dmn$county))

# Compose the stan input
stan_data <- list(y = y, x  = x, u = u, N = N, J = J, county = county)
```

3. Run the model

```
mod <- stan(data = stan_data, file = "lab8_stan_model.stan")
```

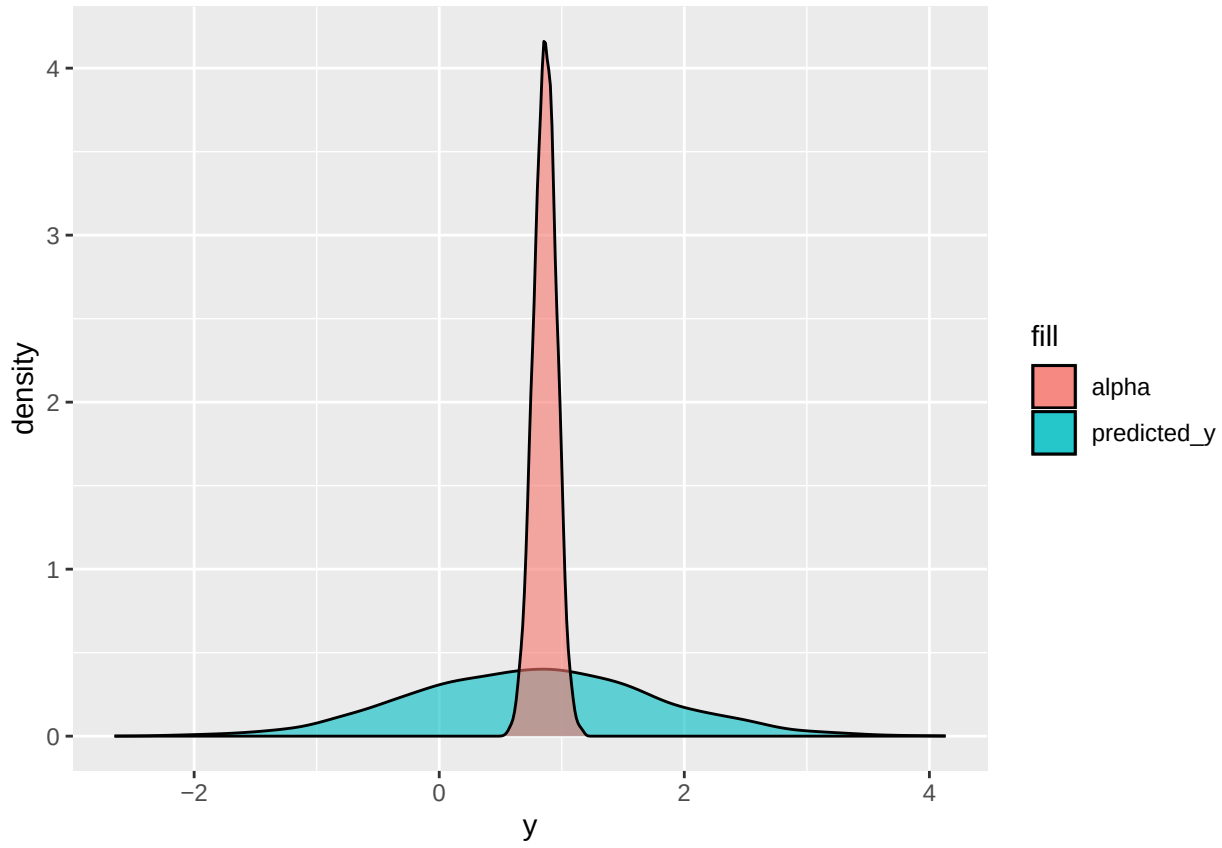
4. For α plot, get median estimates of alpha's, and the 2.5th and 97.5th percentiles. Also get the median (mean fine, easier to pull from summary) of the gamma0 and gamma1. You can then use `geom_abline()` to plot mean regression line.

```
# Extract the results
samps <- extract(mod)

alpha_hat <- apply(samps[["alpha"]], 2, median)
alpha_lower <- apply(samps[["alpha"]], 2, quantile, 0.025)
alpha_upper <- apply(samps[["alpha"]], 2, quantile, 0.975)

alpha_df <- tibble(county = 1:J, median = alpha_hat, lower = alpha_lower, upper = alpha_upper, u = u)
gamma0 <- median(samps[["gamma0"]])
gamma1 <- median(samps[["gamma1"]])

# Provide the plot
ggplot(alpha_df, aes(u, median)) +
  geom_point()+
  geom_errorbar(aes(ymin = lower, ymax = upper))+
  geom_abline(intercept = gamma0, slope = gamma1)
```



5. For the predicted y plot, you will need your posterior predictive samples for y's and then just use `geom_density()`

```
alpha_2 <- samps[["alpha"]][,2]
sigma_y  <- samps[["sigma_y"]]
yrep_2 <- rnorm(alpha_2, sigma_y)

tibble(alpha = alpha_2, y = yrep_2) %>%
  ggplot(aes(y)) +
  geom_density(aes(fill = "predicted_y"), alpha = 0.6)+
  geom_density(aes(alpha, fill = "alpha"), alpha = 0.6)
```


density
4
3
2
1
0
−2
0
2
4
y
fill
alpha
predicted_y