# homework 9

Seongmoon Cho

2022-04-19

```
library(haven)   # for opening dta file
library(ggplot2)  # for plots
library(magrittr)  # for `%>%` operator
library(here)
library(readxl)  # for reading excel files
library(modelsummary)  # for summarizing data
```

```
## Warning: package 'modelsummary' was built under R version 4.1.3
```

```
library(rstan)
rstan_options(auto_write = TRUE)  # save compiled STAN object
options(mc.cores = 2)  # use two cores
library(posterior)
library(bayesplot)
theme_set(theme_classic() +
    theme(panel.grid.major.y = element_line(color = "grey92")))
```

# Research Question

> Do we observe higher PM readings for months in which forest fires were frequently observed?

# Variables

- `pm10` : Particulate matter 10 (micrograms/$m^3$)
- `pm25` : Particulate matter 2.5 (micrograms/$m^3$)
- `myday` : Day of the year in format ddmmyyyy
- `month` : Month of the year
- `rain` : Rainfall (mm)
- `Station_ID` : Pollutant tracking station ID
- `tmp` : Temperature at 20m from ground level (degrees in Celsius)
- `quadrant` : Quadrant in which the station is located (1 = NE, 2 = SE, 3 = SW, 4 = NW)
- `tot` : Total number of forest fires within 1,500km X 1,500km from the center of Bogota, Columbia
- `qtot` : `tot` by quadrant
- `max` : Daily max forest Fire Radiative Power (FRP)
- `qmax` : `tot` by quadrant
- `fire` : A dummy variable indicating months that forest fires were frequently observed

# Import Data

```
data <- read_dta(here("bayes_final1.dta"))
```

# Variable Summary

```
datasummary(daily_avg_pm10  *
                (N + Mean + SD + Min + Max + Histogram) ~
                factor(fire, labels = c("Non_Fire", "Fire")),
            data = data)
```

| | | Non_Fire | Fire |
|---|---|---|---|
| daily_avg_pm10 | N | 11934 | 2301 |
| | Mean | 32.99 | 45.23 |
| | SD | 20.46 | 23.55 |
| | Min | 0.00 | 0.00 |
| | Max | 136.43 | 119.72 |
| | Histogram | _ | _ |

# Model

Let $Y$ = PM10, $G$ = Forest Fire

Model:

$$Y_{i,G=0} \sim N(\mu_1, \sigma_1)$$
$$Y_{i,G=1} \sim N(\mu_2, \sigma_2)$$

Prior:

$$\mu_1 \sim N(36, 20)$$
$$\mu_2 \sim N(45, 23)$$
$$\sigma_1 \sim N^+(0, 2)$$
$$\sigma_2 \sim N^+(0, 2)$$

# Running Stan

We used 4 chains, each with 4,000 iterations (first 2,000 as warm-ups).

```
# 1. form the data list for Stan
stan_dat <- with(data,
    list(N1 = sum(fire == 0),
         N2 = sum(fire == 1),
         y1 = daily_avg_pm10[which(fire == 0)],
         y2 = daily_avg_pm10[which(fire == 1)])
)
# 2. Run Stan
m1 <- stan(
    file = here("normal_2group.stan"),
    data = stan_dat,
    seed = 20220419,  # for reproducibility
    iter = 4000
)
```

```
## Warning in readLines(file, warn = TRUE): incomplete final line
## found on 'C:\SMC\USC_PhD\PhD_Courses\PSYC573_Bayesian Data
## Analysis\final\normal_2group.stan'
```

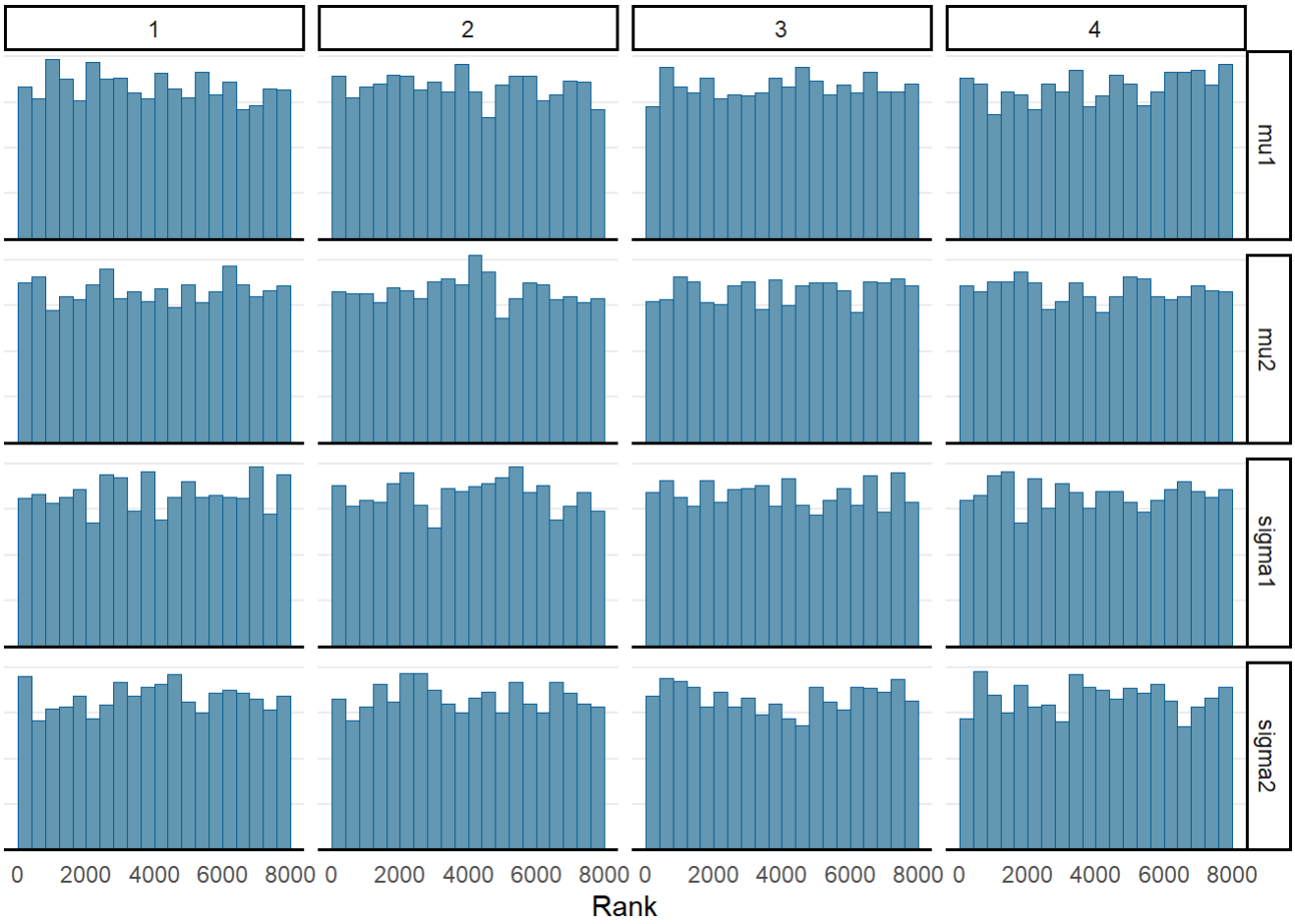
# Results

```
print(m1, pars = c("mu1", "mu2", "sigma1", "sigma2"))
```

```
## Inference for Stan model: normal_2group.
## 4 chains, each with iter=4000; warmup=2000; thin=1;
## post-warmup draws per chain=2000, total post-warmup draws=8000.
##
##         mean se_mean   sd  2.5%   25%   50%   75% 97.5% n_eff Rhat
## mu1    31.92    0.00 0.18 31.57 31.80 31.92 32.04 32.27  7931    1
## mu2    37.14    0.01 0.46 36.24 36.83 37.15 37.46 38.02  7725    1
## sigma1 20.15    0.00 0.13 19.90 20.06 20.15 20.23 20.40  7841    1
## sigma2 22.54    0.00 0.30 21.96 22.34 22.54 22.75 23.14  7596    1
##
## Samples were drawn using NUTS(diag_e) at Tue Apr 19 23:00:17 2022.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

As shown in the graph below, the chains mixed well.

```
mcmc_rank_hist(m1, pars = c("mu1", "mu2", "sigma1", "sigma2"))
```

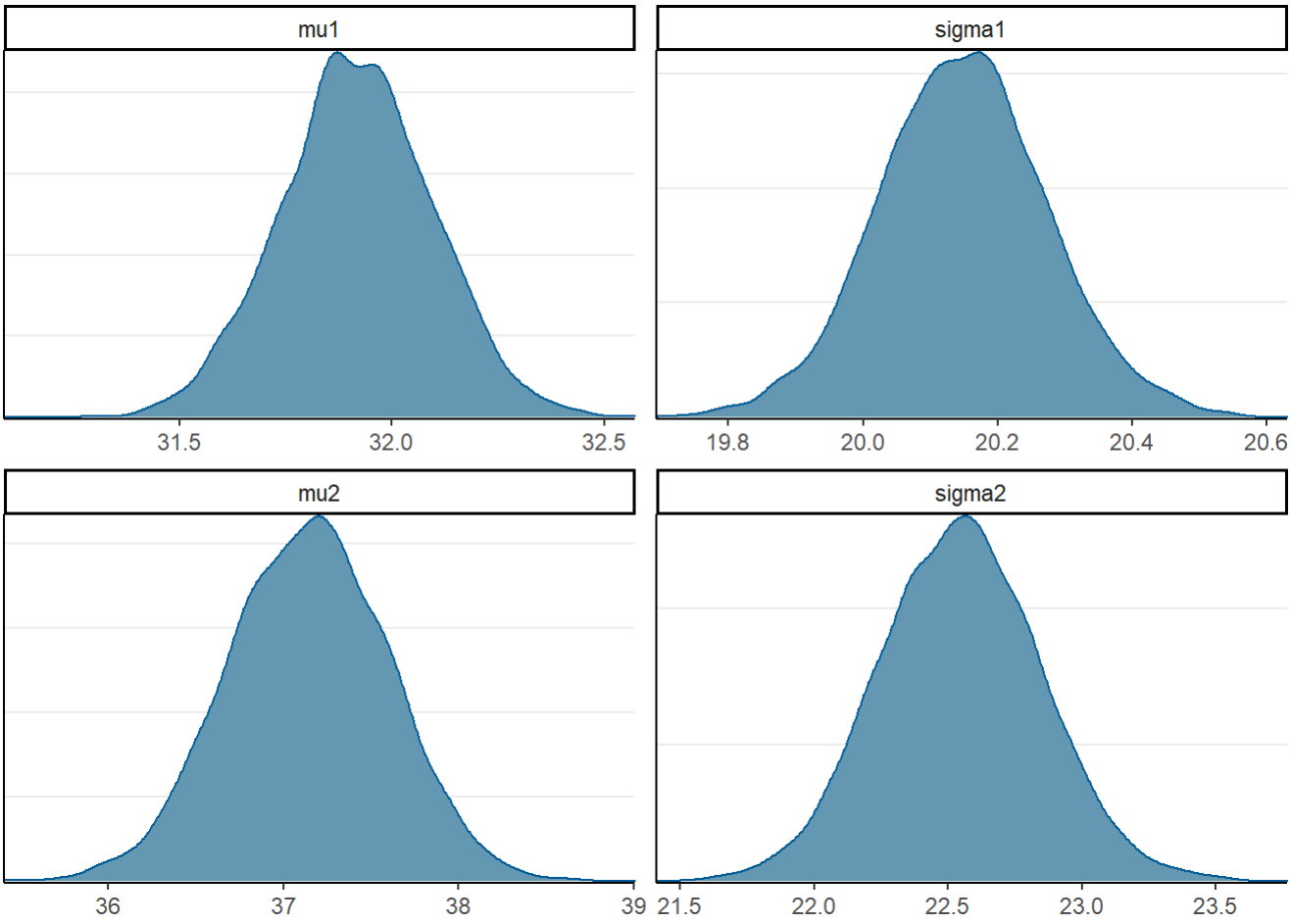



The following table shows the posterior distributions of $\mu_1$, $\mu_2$, $\sigma_1$, $\sigma_2$, and $\mu_2 - \mu_1$.

```
summ_m1 <- as_draws_df(m1) %>%
    subset_draws(variable = c("mu1", "mu2", "sigma1", "sigma2")) %>%
    mutate_variables(`mu2 - mu1` = mu2 - mu1) %>%
    summarise_draws()
knitr::kable(summ_m1, digits = 2)
```

| variable | mean | median | sd | mad | q5 | q95 | rhat | ess_bulk | ess_tail |
|---|---|---|---|---|---|---|---|---|---|
| mu1 | 31.92 | 31.92 | 0.18 | 0.17 | 31.62 | 32.21 | 1 | 7937.92 | 6303.37 |
| mu2 | 37.14 | 37.15 | 0.46 | 0.47 | 36.39 | 37.89 | 1 | 7738.86 | 6615.03 |
| sigma1 | 20.15 | 20.15 | 0.13 | 0.12 | 19.94 | 20.36 | 1 | 7866.02 | 5649.64 |
| sigma2 | 22.54 | 22.54 | 0.30 | 0.30 | 22.06 | 23.04 | 1 | 7634.57 | 6098.18 |
| mu2 - mu1 | 5.22 | 5.22 | 0.49 | 0.49 | 4.41 | 6.02 | 1 | 7747.00 | 6239.31 |

```
mcmc_dens(m1,
          pars = c("mu1","sigma1","mu2","sigma2"))
```




The analysis showed that on average, months with frequent forest fire showed higher PM concentrate than months that did not observe frequent forest fire, with a posterior mean of 5.22 and a 90% CI of [4.41, 6.02].