Week 9: Hierarchical GLM

18/03/24

Lip cancer

Here is the lip cancer data that was used in the lecture.

- `aff.i` is proportion of male population working outside in each region
- `observe.i` is observed deaths in each region
- `expect.i` is expected deaths, based on region-specific age distribution and national-level age-specific mortality rates.

```
observe.i <- c(
  5,13,18,5,10,18,29,10,15,22,4,11,10,22,13,14,17,21,25,6,11,21,13,5,19,18,14,17,3,10,
  7,3,12,11,6,16,13,6,9,10,4,9,11,12,23,18,12,7,13,12,12,13,6,14,7,18,13,9,6,8,7,6,16,4,6,
  17,5,7,2,9,7,6,12,13,17,5,5,6,12,10,16,10,16,15,18,6,12,6,8,33,15,14,18,25,14,2,73,13,14
  12,10,3,11,3,11,13,11,13,10,5,18,10,23,5,9,2,11,9,11,6,11,5,19,15,4,8,9,6,4,4,2,12,12,11
  8,12,11,23,7,16,46,9,18,12,13,14,14,3,9,15,6,13,13,12,8,11,5,9,8,22,9,2,10,6,10,12,9,11,
  9,11,11,0,9,3,11,11,11,5,4,8,9,30,110)
expect.i <- c(
    6.17,8.44,7.23,5.62,4.18,29.35,11.79,12.35,7.28,9.40,3.77,3.41,8.70,9.57,8.18,4.35,
    4.91,10.66,16.99,2.94,3.07,5.50,6.47,4.85,9.85,6.95,5.74,5.70,2.22,3.46,4.40,4.05,5.74
    16.99,6.19,5.56,11.69,4.69,6.25,10.84,8.40,13.19,9.25,16.98,8.39,2.86,9.70,12.12,12.94
    10.34,5.09,3.29,17.19,5.42,11.39,8.33,4.97,7.14,6.74,17.01,5.80,4.84,12.00,4.50,4.39,1
    6.42,5.26,4.59,11.86,4.05,5.48,13.13,8.72,2.87,2.13,4.48,5.85,6.67,6.11,5.78,12.31,10.
    2.52,6.22,14.29,5.71,37.93,7.81,9.86,11.61,18.52,12.28,5.41,61.96,8.55,12.07,4.29,19.4
    12.90,4.76,5.56,11.11,4.76,10.48,13.13,12.94,14.61,9.26,6.94,16.82,33.49,20.91,5.32,6.
    12.94,16.07,8.87,7.79,14.60,5.10,24.42,17.78,4.04,7.84,9.89,8.45,5.06,4.49,6.25,9.16,1
    9.57,5.83,9.21,9.64,9.09,12.94,17.42,10.29,7.14,92.50,14.29,15.61,6.00,8.55,15.22,18.4
    18.37,13.16,7.69,14.61,15.85,12.77,7.41,14.86,6.94,5.66,9.88,102.16,7.63,5.13,7.58,8.0
    18.75,12.33,5.88,64.64,8.62,12.09,11.11,14.10,10.48,7.00,10.23,6.82,15.71,9.65,8.59,8.
    12.31,8.91,50.10,288.00)
aff.i <- c(0.2415,0.2309,0.3999,0.2977,0.3264,0.3346,0.4150,0.4202,0.1023,0.1752,
        0.2548,0.3248,0.2287,0.2520,0.2058,0.2785,0.2528,0.1847,0.3736,0.2411,
        0.3700,0.2997,0.2883,0.2427,0.3782,0.1865,0.2633,0.2978,0.3541,0.4176,
        0.2910,0.3431,0.1168,0.2195,0.2911,0.4297,0.2119,0.2698,0.0874,0.3204,
        0.1839,0.1796,0.2471,0.2016,0.1560,0.3162,0.0732,0.1490,0.2283,0.1187,
        0.3500,0.2915,0.1339,0.0995,0.2355,0.2392,0.0877,0.3571,0.1014,0.0363,
        0.1665,0.1226,0.2186,0.1279,0.0842,0.0733,0.0377,0.2216,0.3062,0.0310,
        0.0755,0.0583,0.2546,0.2933,0.1682,0.2518,0.1971,0.1473,0.2311,0.2471,
        0.3063,0.1526,0.1487,0.3537,0.2753,0.0849,0.1013,0.1622,0.1267,0.2376,
        0.0737,0.2755,0.0152,0.1415,0.1344,0.1058,0.0545,0.1047,0.1335,0.3134,
        0.1326,0.1222,0.1992,0.0620,0.1313,0.0848,0.2687,0.1396,0.1234,0.0997,
        0.0694,0.1022,0.0779,0.0253,0.1012,0.0999,0.0828,0.2950,0.0778,0.1388,
        0.2449,0.0978,0.1144,0.1038,0.1613,0.1921,0.2714,0.1467,0.1783,0.1790,
        0.1482,0.1383,0.0805,0.0619,0.1934,0.1315,0.1050,0.0702,0.1002,0.1445,
        0.0353,0.0400,0.1385,0.0491,0.0520,0.0640,0.1017,0.0837,0.1462,0.0958,
        0.0745,0.2942,0.2278,0.1347,0.0907,0.1238,0.1773,0.0623,0.0742,0.1003,
        0.0590,0.0719,0.0652,0.1687,0.1199,0.1768,0.1638,0.1360,0.0832,0.2174,
        0.1662,0.2023,0.1319,0.0526,0.0287,0.0405,0.1616,0.0730,0.1005,0.0743,
        0.0577,0.0481,0.1002,0.0433,0.0838,0.1124,0.2265,0.0436,0.1402,0.0313,
        0.0359,0.0696,0.0618,0.0932,0.0097)
```

Question 1

Explain a bit more what the `expect.i` variable is. For example, if a particular area has an expected deaths of 16, what does this mean?

The `expect.i` variable represents the expected deaths from lip cancer in each region, given the regions specific age distribution and the national-level age-specific mortality rates. The `expected.i` is not the actual observable death, but is the baseline number against which the `observe.i` can be compared.

In order to calculate `epect.i`, the whole reference populaton was split-up into age deciles. In each decile, the observed overall prevalence of lip cancer mortality among males was computed. The overall prevalence in each age decile was then multiplied with the population size of each region in the respective age decile. The sum of the products, i.e. `expect/i`, is an estimate of the true count in that region if the risk of dying from lip cancer for a male were equal across all regions in the reference population.

So if a particular area has an expected deaths of 16, it means that given the distribution of the population woking outside in that area, and considering the average mortality rates due to lip cancer at the national level for those age groups, we would expect, on average, 16 deaths to occur in that regions due to lio cancer in the given time frame.

```
library(tidyverse)
library(here)
library(rstan)
library(tidybayes)
library(loo)

data<-data.frame(aff.i,observe.i,expect.i)
data<-data|>mutate(SMR.i=observe.i/expect.i,aff.i.centered=aff.i-mean(aff.i))
head(data)
```

```
   aff.i observe.i expect.i     SMR.i aff.i.centered
1 0.2415         5     6.17 0.8103728      0.0746559
2 0.2309        13     8.44 1.5402844      0.0640559
3 0.3999        18     7.23 2.4896266      0.2330559
4 0.2977         5     5.62 0.8896797      0.1308559
5 0.3264        10     4.18 2.3923445      0.1595559
6 0.3346        18    29.35 0.6132879      0.1677559
```

Question 2

Run four different models in Stan with three different set-ups for estimating θ_i, that is the relative risk of lip cancer in each region:

1. Intercept α_i is same in each region $= \alpha$

```
my_data <- list(
  N = length(observe.i),
  x = aff.i-mean(aff.i),
  obs_death = observe.i,
  exp_death = expect.i
)

# fit1<-rstan::stan(data = data,
#               file = here("labs/lip_cancer1.stan"),
#               iter = 1000,
#               seed = 243,
#               chains = 3)
# save(fit1, file = "lab9fit1.rda")
load("lab9fit1.rda")
summary(fit1)$summary[c("alpha","beta"),]
```

```
              mean      se_mean         sd        2.5%         25%          50%
alpha -0.008997547 0.0005779209 0.02154681 -0.05197262 -0.02414655 -0.008135722
beta   2.422092597 0.0056574038 0.17856188  2.06841480  2.30180787  2.422264326
             75%      97.5%     n_eff      Rhat
alpha 0.00550633 0.03186939 1390.0459 0.9992492
beta  2.54805412 2.76870056  996.1922 1.0017156
```

2. Intercept α_i is different in each region and modeled separately

```
# fit2<-rstan::stan(data = data,
#               file = here("labs/lip_cancer2.stan"),
#               iter = 1000,
#               seed = 243,
#               chains = 3)
# save(fit2, file = "lab9fit2.rda")

load("lab9fit2.rda")
head(summary(fit2)$summary)
```

```
               mean     se_mean        sd        2.5%        25%        50%
alpha[1] -0.3435784 0.007987945 0.4032565 -1.19301207 -0.6021779 -0.3281612
alpha[2]  0.2789236 0.004940352 0.2679736 -0.28801272  0.1111505  0.2947627
alpha[3]  0.5020027 0.009041295 0.2756205 -0.03859371  0.3150936  0.5143341
alpha[4] -0.3332257 0.009425279 0.4238050 -1.23905517 -0.5992107 -0.3104127
alpha[5]  0.5102656 0.007452085 0.3343101 -0.16574723  0.2801271  0.5222797
alpha[6] -0.7417356 0.007090717 0.2424020 -1.24273580 -0.8980748 -0.7280608
                 75%      97.5%     n_eff      Rhat
alpha[1] -0.05969975  0.3672394 2548.5476 0.9990731
alpha[2]  0.45327443  0.7694537 2942.1726 0.9993853
alpha[3]  0.69318609  1.0021003  929.3123 1.0036734
alpha[4] -0.04725356  0.4619020 2021.8258 1.0006697
alpha[5]  0.74586676  1.1469244 2012.5353 0.9998688
alpha[6] -0.58320094 -0.2900966 1168.6710 1.0016372
```

3. Intercept α_i is different in each region and the intercept is modeled hierarchically

```
# fit3<-rstan::stan(data = data,
#               file = here("labs/lip_cancer3.stan"),
#               iter = 1000,
#               seed = 243,
#               chains = 3)
# save(fit3, file = "lab9fit3.rda")
```

```
load("lab9fit3.rda")
head(summary(fit3)$summary)
```

```
                mean      se_mean         sd        2.5%         25%
mu        0.08615442 0.0008658295 0.03503414  0.01877793  0.06218049
sigma     0.38494315 0.0010211417 0.03071727  0.32924461  0.36392123
alpha[1] -0.14449158 0.0046154304 0.27724460 -0.68515351 -0.32869961
alpha[2]  0.20339818 0.0036535003 0.22501800 -0.24361741  0.04992053
alpha[3]  0.29295956 0.0043525143 0.22440754 -0.15212351  0.14229895
alpha[4] -0.15488639 0.0049281664 0.28438994 -0.73832027 -0.34826299
                 50%        75%     97.5%     n_eff      Rhat
mu        0.08611062 0.11015105 0.1530558 1637.2623 0.9986522
sigma     0.38381052 0.40468646 0.4516836  904.8845 1.0026868
alpha[1] -0.13879927 0.04543891 0.3828400 3608.2930 0.9986690
alpha[2]  0.20149041 0.35762289 0.6371989 3793.2916 0.9993600
alpha[3]  0.30093382 0.44733280 0.7075024 2658.2418 0.9985764
alpha[4] -0.15307620 0.04892763 0.3715386 3330.1035 0.9985637
```

Note in all three cases, use the proportion of male population working outside in each region as a covariate.

Question 3

Make two plots (appropriately labeled and described) that illustrate the differences in estimated θ_i's across regions and the differences in θs across models.

The plot below shows the distribution of posterior estimates of theta across different models. We observe that the theta estimates for model1,mode2 and model3 are -0.009,0.046 and 0.086 respectively. In addition, the standard deviation for model 1's theta estimates is the smallest, with a value of 0.24, whereas the standard deviation estiamted from model 2 is the largest, with a value of 0.599.

```
estimated_thetas1 = rstan::extract(fit1)$theta
estimated_thetas2 = rstan::extract(fit2)$theta
estimated_thetas3 = rstan::extract(fit3)$theta
mean(estimated_thetas1)
```

```
[1] -0.008997547
```

```
mean(estimated_thetas2)
```

```
[1] 0.0460639
```

```
mean(estimated_thetas3)
```

```
[1] 0.08618693
```

```
sd(estimated_thetas1)
```

```
[1] 0.2416157
```

```
sd(estimated_thetas2)
```

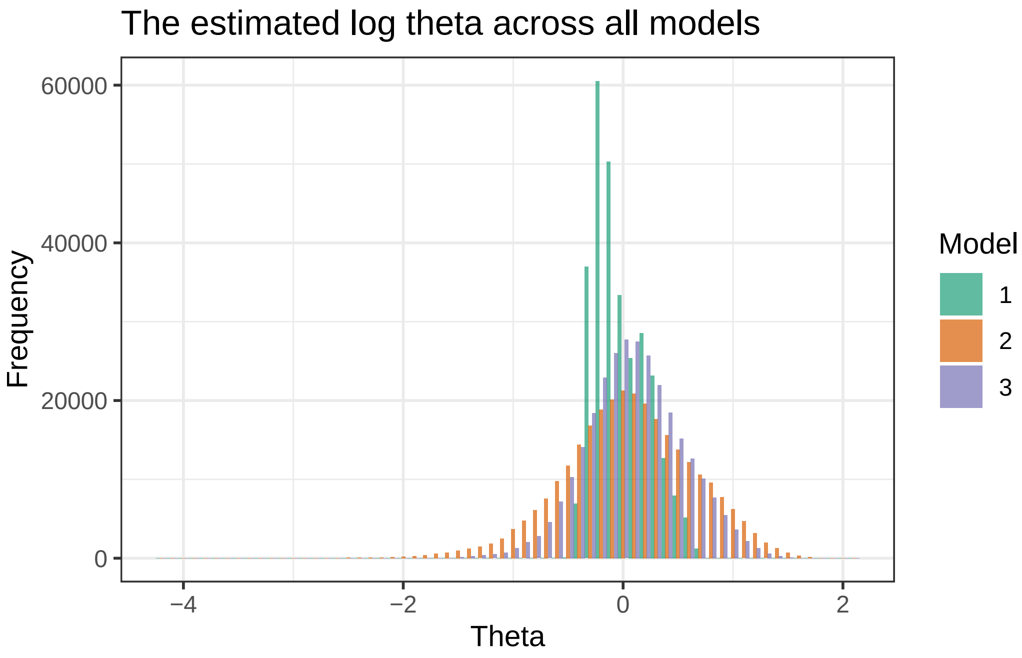
```
[1] 0.5989602
```

```
sd(estimated_thetas3)
```

```
[1] 0.4417448
```

```
a = c(estimated_thetas1, estimated_thetas2, estimated_thetas3)
mod = c(rep(1, length(estimated_thetas1)), rep(2, length(estimated_thetas2)), rep(3, lengt
df = data.frame(theta = a, model = factor(mod)) # Ensure 'model' is a factor for better pl

ggplot(df, aes(x = theta, fill = model)) +
  geom_histogram(position = "dodge", binwidth = 0.1, alpha = 0.7) + # Adjust binwidth as n
  labs(x = "Theta", y = "Frequency", fill = "Model",title="The estimated log theta across
  theme_bw() +
  scale_fill_brewer(palette = "Dark2")
```

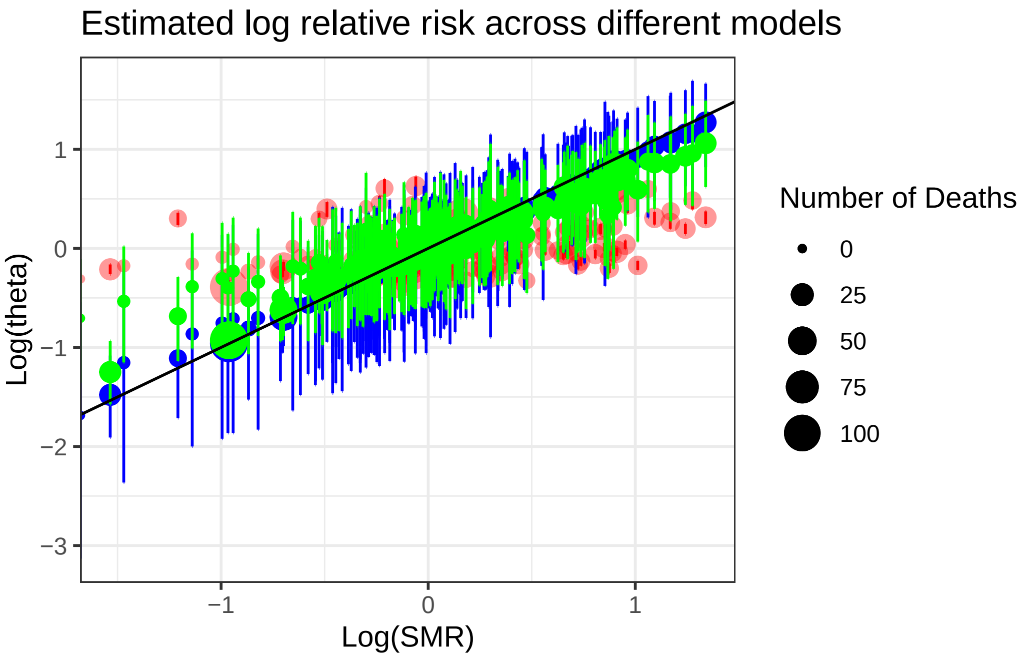


We then plot the log theta of three different models against the log_SMR. The red, green and blue represents model1, model2, and model3, respectively. We observe that model1 is not capturing the general trend very well, wehreas model3 captures the general trend the best.

```
library(posterior)
theta_estimates <- function(model, median_name, lower_name, upper_name) {
  model |>
  posterior::as_draws_df() |>
tidybayes::gather_draws(theta[i]) |>
tidybayes::median_qi() |>
rename_with(~c(median_name, lower_name, upper_name), .cols = c(".value", ".lower", ".upper
select(i, starts_with("median"), starts_with("lower"), starts_with("upper"))
}
fit_estimate1 <- theta_estimates(fit1, "median_mod1", "lower_mod1", "upper_mod1")
fit_estimate2 <- theta_estimates(fit2, "median_mod2", "lower_mod2", "upper_mod2")
fit_estimate3 <- theta_estimates(fit3, "median_mod3", "lower_mod3", "upper_mod3")
```

```
res1 <- list(fit_estimate1, fit_estimate2, fit_estimate3) %>%
reduce(left_join, by = "i")
long_data <- res1 |>
pivot_longer(cols = starts_with('median'),
names_to = 'model',
values_to = 'theta') |>
mutate(model = str_remove(model, 'median_'))
```

```
res1 |>
mutate(
deaths = observe.i,
log_smr = log(observe.i / expect.i)
) |>
ggplot(aes(log_smr, median_mod1, color = "red")) +
geom_point(aes(size = deaths),alpha = 0.4) +
geom_errorbar(aes(ymin = lower_mod1, ymax = upper_mod1, color = "red")) +
geom_point(aes(log_smr, median_mod2, color = "blue", size = deaths)) +
geom_errorbar(aes(ymin = lower_mod2, ymax = upper_mod2, color = "blue"))+
geom_point(aes(log_smr, median_mod3, color = "green", size = deaths)) +
geom_errorbar(aes(ymin = lower_mod3, ymax = upper_mod3, color = "green")) +
geom_abline(slope = 1, intercept = 0) +
labs(
title = "Estimated log relative risk across different models",
x = "Log(SMR)",
y = "Log(theta)",
size = "Number of Deaths"
) +
scale_color_identity() +
theme_bw()
```



Question 4

Using tool of your choice, decide which model is the best, and justify your choice.

We compare the models using epld using `loo` package. From the difference function, we find that model3 has the largest elpd, indicating that in general, our third model has the highest predictive performance.

```
loglik1 <- rstan::extract(fit1)[["log_lik"]]
loo1 <- loo::loo(loglik1, save_psis = TRUE)
loglik2 <- rstan::extract(fit2)[["log_lik"]]
loo2 <- loo::loo(loglik2, save_psis = TRUE)
loglik3 <- rstan::extract(fit3)[["log_lik"]]
loo3 <- loo::loo(loglik3, save_psis = TRUE)
```

```
loo_compare(loo1,loo2,loo3)
```

```
       elpd_diff se_diff
model3    0.0       0.0 
model2  -16.5       7.9 
model1 -154.0      45.6 
```