Hierarchical GLM

Monica Alexander

March 10 2022

Please hand in Rmd, pdf, and stan files. Due next Wednesday because of delay in lecture.

Lip cancer

Here is the lip cancer data as seen in the lecture.

- `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,
  7,3,12,11,6,16,13,6,9,10,4,9,11,12,23,18,12,7,13,12,12,13,6,14,7,18,13,9,6,8,7,6,16,4,6,12,5,5,
  17,5,7,2,9,7,6,12,13,17,5,5,6,12,10,16,10,16,15,18,6,12,6,8,33,15,14,18,25,14,2,73,13,14,6,20,8,
  12,10,3,11,3,11,13,11,13,10,5,18,10,23,5,9,2,11,9,11,6,11,5,19,15,4,8,9,6,4,4,2,12,12,11,9,7,7,
  8,12,11,23,7,16,46,9,18,12,13,14,14,3,9,15,6,13,13,12,8,11,5,9,8,22,9,2,10,6,10,12,9,11,32,5,11,
  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,
    4.91,10.66,16.99,2.94,3.07,5.50,6.47,4.85,9.85,6.95,5.74,5.70,2.22,3.46,4.40,4.05,5.74,6.36,5.13,
    16.99,6.19,5.56,11.69,4.69,6.25,10.84,8.40,13.19,9.25,16.98,8.39,2.86,9.70,12.12,12.94,9.77,
    10.34,5.09,3.29,17.19,5.42,11.39,8.33,4.97,7.14,6.74,17.01,5.80,4.84,12.00,4.50,4.39,16.35,6.02,
    6.42,5.26,4.59,11.86,4.05,5.48,13.13,8.72,2.87,2.13,4.48,5.85,6.67,6.11,5.78,12.31,10.56,10.23,
    2.52,6.22,14.29,5.71,37.93,7.81,9.86,11.61,18.52,12.28,5.41,61.96,8.55,12.07,4.29,19.42,8.25,
    12.90,4.76,5.56,11.11,4.76,10.48,13.13,12.94,14.61,9.26,6.94,16.82,33.49,20.91,5.32,6.77,8.70,
    12.94,16.07,8.87,7.79,14.60,5.10,24.42,17.78,4.04,7.84,9.89,8.45,5.06,4.49,6.25,9.16,12.37,8.40,
    9.57,5.83,9.21,9.64,9.09,12.94,17.42,10.29,7.14,92.50,14.29,15.61,6.00,8.55,15.22,18.42,5.77,
    18.37,13.16,7.69,14.61,15.85,12.77,7.41,14.86,6.94,5.66,9.88,102.16,7.63,5.13,7.58,8.00,12.82,
    18.75,12.33,5.88,64.64,8.62,12.09,11.11,14.10,10.48,7.00,10.23,6.82,15.71,9.65,8.59,8.33,6.06,
    12.31,8.91,50.10,288.00)

stan_data <- list(N = length(observe.i),
                  y = observe.i,
                  e = expect.i)
```

Question 1

Explain a bit more what the `expect.i` variable is. For example, if a particular area has an expected deaths of 6, what does this mean?

Cancer is more likely to occur among eldery population. If we have expected region specific level of 6 it means that the region has relatively young population and we do not expect many cases of lip cancer here.

Question 2

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

```
stan_data <- list(N = length(observe.i),
                  y = observe.i,
                  e = expect.i)
```

1. θ_i is same in each region $= \theta$

```
# mod1 <- stan(data = stan_data,
#              file = "lab8_mod1.stan",
#              iter = 500,
#              seed = 161198
#              )
```

```
#saveRDS(mod1, "lab8_mod1_new.rds")
mod1 <- readRDS("lab8_mod1_new.rds")
```

2. θ_i is different in each region and modeled separately

```
# mod2 <- stan(data = stan_data,
#              file = "lab8_mod2.stan",
#              iter = 500,
#              seed = 161198
#              )
```

```
#saveRDS(mod2, "lab8_mod2_new.rds")
mod2 <- readRDS("lab8_mod2_new.rds")
```

3. θ_i is different in each region and modeled hierarchically

```
# mod3 <- stan(data = stan_data,
#              file = "lab8_mod3.stan",
#              iter = 500,
#              seed = 161198
#              )
```

```
#saveRDS(mod3, "lab8_mod3_new.rds")
mod3 <- readRDS("lab8_mod3_new.rds")
```

Question 3

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

```
mod1_thetas = summary(mod1)$summary[c('theta'), c('mean')]
mod2_thetas = summary(mod2)$summary[1:100, c('mean')]
mod3_thetas = summary(mod3)$summary[1:100, c('mean')]
```

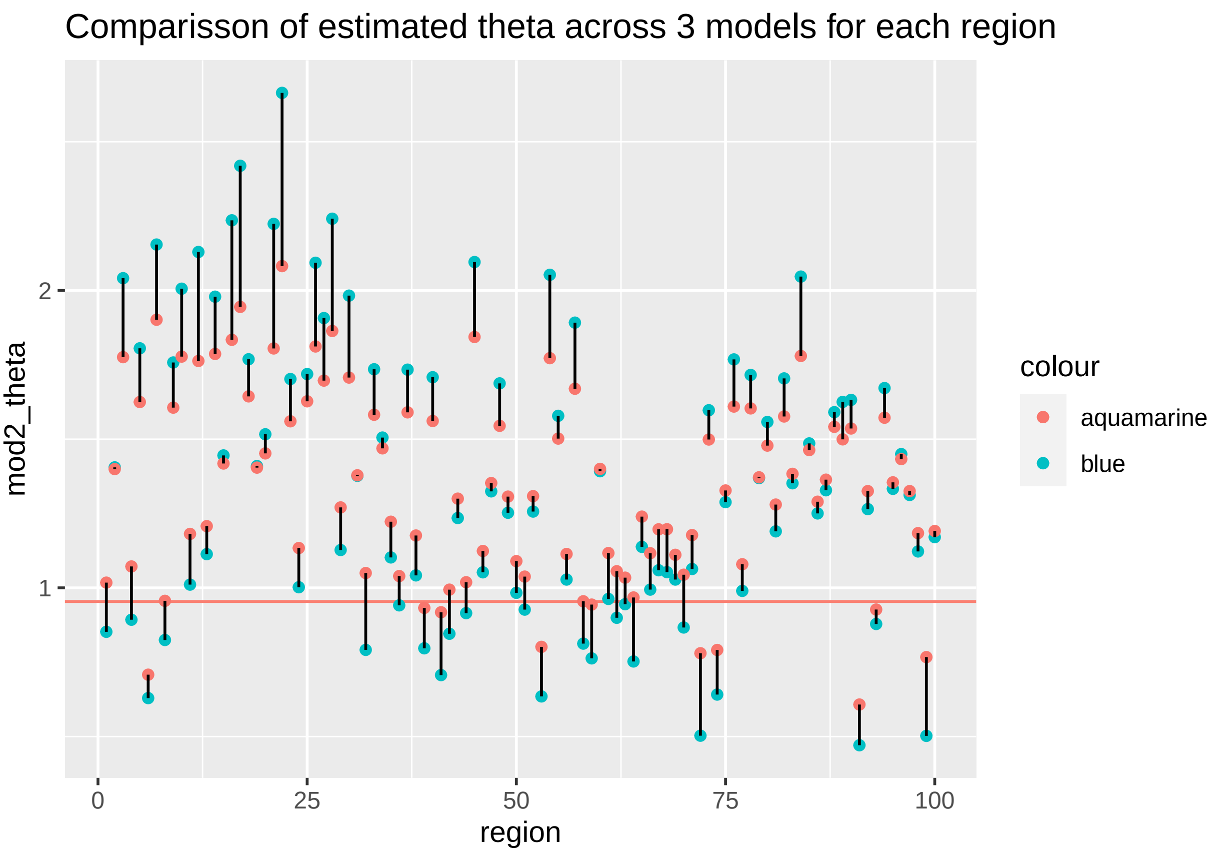
```
region <- seq(1,100,1)
region
```

```
##   [1]   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18
##  [19]  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36
##  [37]  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54
##  [55]  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72
##  [73]  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90
##  [91]  91  92  93  94  95  96  97  98  99 100
```

```
tdf <- data.frame(region)
```

```
tdf$mod1_theta <- mod1_thetas
tdf$mod2_theta <- mod2_thetas
tdf$mod3_theta <- mod3_thetas
```

```
ggplot(data = tdf)+
  geom_hline(yintercept = tdf$mod1_theta, color = 'salmon') +
  geom_point(aes(x = region, y = mod2_theta, color = 'blue'))+
  geom_point(aes(x = region, y = mod3_theta, color = 'aquamarine'))+
  geom_segment(aes(x = region,
                   y = mod2_theta,
                   xend =region,
                   yend = mod3_theta))+
  ggtitle('Comparisson of estimated theta across 3 models for each region')
```



Question 4

Rerun model 3 (the hierarchical model), but also including an overdispersion parameter. Compare the two models and decide which is more appropriate.

```
# mod4 <- stan(data = stan_data,
#              file = "lab8_mod4.stan",
#              iter = 500,
#              seed = 161198
#              )
```

```
#saveRDS(mod4, "lab8_mod4_new.rds")
mod4 <- readRDS("lab8_mod4_new.rds")
```

```
loglik3 <- rstan::extract(mod3)[["log_lik"]]
loglik4 <- rstan::extract(mod4)[["log_lik"]]
```

```
loo3 = loo(loglik3, save_psis = T )
loo4 = loo(loglik4, save_psis = T )
```

```
loo_compare(loo3, loo4)
```

```
##        elpd_diff se_diff
## model2  0.0       0.0
## model1 -8.1       8.3
```