Week 9: Hierarchical GLM

Priyanka Verma

20/03/23

Lip cancer

Here is the lip cancer data given to you in terribly unreproducible and error-prone format.

- `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.

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?

Answer 1

Expected deaths is the implied number of lip cancer deaths for a particular region given that the region's age structure and the national level age-specific mortality rates for lip cancer. For example, an expected number of deaths of 6 would mean that for that particular region, we would expect 6 lip cancer deaths if this region were to experience the same age specific mortality rates as at the national level.

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:

1. Intercept α_i is same in each region $= \alpha$
2. α_i is different in each region and modeled separately (with covariate)
3. α_i is different in each region and the intercept is modeled hierarchically (with covariate)

$$y_i|\theta_i \sim Posson(\theta_i.e_i)$$

$$\log \theta_i = \alpha_i + \beta x_i$$

$$\alpha_i \sim N(0, 1)$$

- These have been trained in the code and output has not been printed here in the pdf.

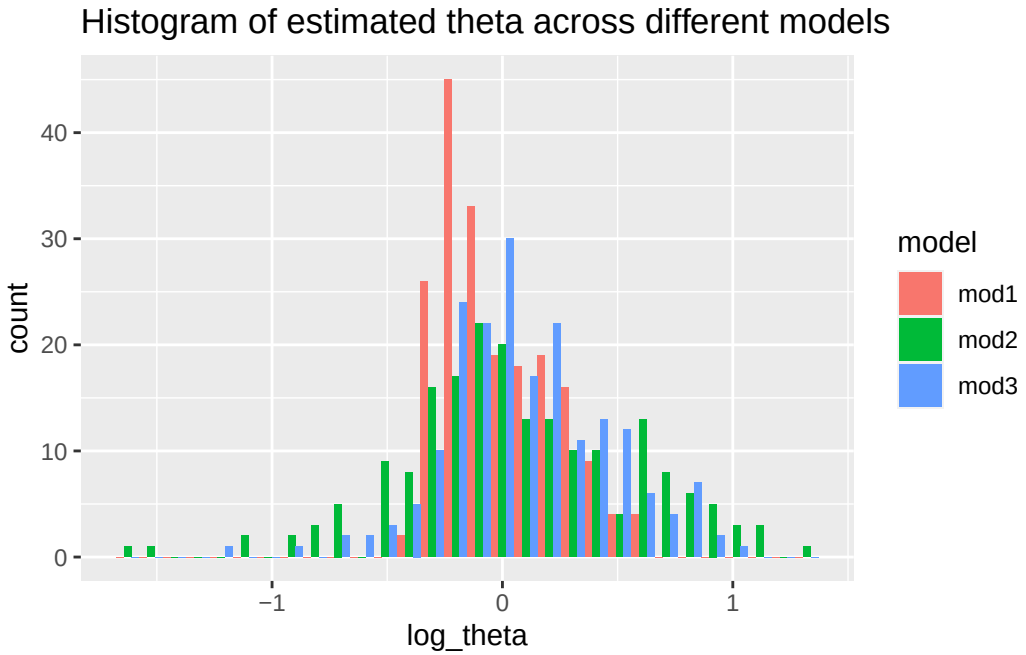
MODEL 2

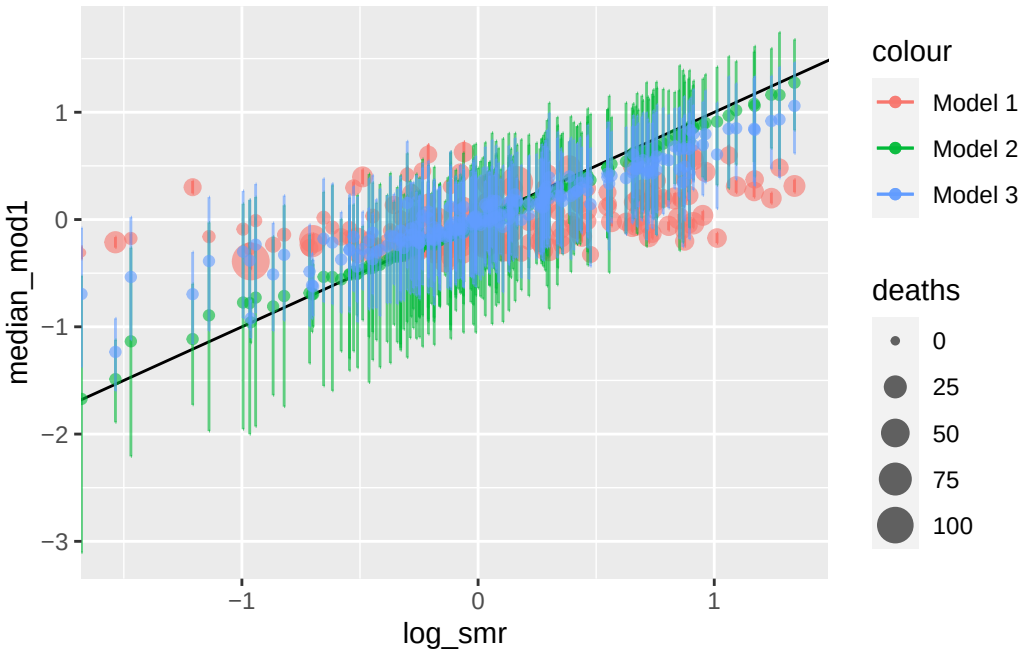
MODEL 3

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 that model 1 has the least variation, while model 2 and model 3 do not differ much.





Model 1 is not performing good based on the observed data as the intercept is fixed. Model 2 and 3 perform similar, but model 2 has more uncertainty than model 3, because we are predicting the mean in each region and there is polling of information, therefore there is a bias variance tradeoff. On the contrary, estimates are shrunk to some global mean in model 3 due

to which the uncertainty is reduced in model 3.