SCI 2025: Homework 9

Setup

This homework is based on practice problem 13H1 from *Statistical Rethinking*.

In 1980, a typical Bengali woman could have 5 or more children in her lifetime. By the year 2000, a typical Bengali woman had only 2 or 3. You're going to look at a historical set of data, when contraception was widely available but many families chose not to use it. These data reside in `data(bangladesh)` and come from the 1988 Bangladesh Fertility Survey. Each row is one of 1934 women. There are six variables, but you can focus on two of them for this practice problem:

(1) `district`: ID number of administrative district each woman resided in
(2) `use.contraception`: An indicator (0/1) of whether the woman was using contraception

The first thing to do is ensure that the cluster variable, `district`, is a contiguous set of integers. Recall that these values will be index values inside the model. If there are gaps, you'll have parameters for which there is no data to inform them. Worse, the model probably won't run. Look at the unique values of the `district` variable:

```
library(rethinking)
data(bangladesh)
```

```
d <- bangladesh
sort(unique(d$district))
```

```
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
[26] 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
[51] 51 52 53 55 56 57 58 59 60 61
```

District 54 is absent. So district isn't yet a good index variable, because it's not contiguous. This is easy to fix. Just make a new variable that is contiguous. This is enough to do it:

```
d$district_id <- as.integer(as.factor(d$district))
sort(unique(d$district_id))
```

```
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
[26] 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
[51] 51 52 53 54 55 56 57 58 59 60
```

Now there are 60 values, contiguous integers 1 to 60.

Question 1:

Now, focus on predicting `use.contraception`, clustered by `district_id`. Fit both:

(1) a traditional fixed-effects model that uses an index variable for district and (2) a multilevel model with varying intercepts for district.

Plot the predicted proportions of women in each district using contraception, for both the fixed-effects model and the varying-effects model. That is, make a plot in which district ID is on the horizontal axis and expected proportion using contraception is on the vertical. Make one plot for each model, or layer them on the same plot, as you prefer. How do the models disagree? Can you explain the pattern of disagreement?

Question 2:

Perform model of the fixed and varying effects models using PSIS-LOO. What do you conclude about their out-of-sample predictive accuracy? Are there any problematic districts flagged by the pareto-k values?

Question 3:

Fit a 3rd model that, instead of Gaussian-distributed varying effects, uses student-t distributed. This means including an additional degree of freedom parameter (`nu`) in the model, that you may either fix to a single value or estimate from the data. Then, compare both the predictions and PSIS-LOO values of the 3 models.

Question 4:

Building upon your preferred model from the previous question, use the indicator variable of `urban` as a predictor of `use.contraception` in addition to district. How does the inclusion of this predictor change the magnitude of the district-level fixed or varying effects? Can you propose an explanation for why this might be?