Practicum on Introduction to Bayesian Analysis

Alexandra M. Schmidt and Carlo Zaccardi

03-06, June 2025

Question 1

The distribution of radon in American houses varies greatly. Some houses have dangerously high concentrations. The EPA did a study of 80,000 houses throughout the country, in order to better understand the distribution of radon. Two important predictors were available:

- whether the measurement was taken in the basement, or the first floor, and
- the level of uranium in the county. Higher levels of uranium are expected to lead to higher radon levels, in general. And, in general, more radon will be measured in the basement than on the first floor.

The file `radon.txt` contains information about measurements made at houses in the state of Minnesota. There are four variables available:

- floor: Indicator for radon measurement made on the first floor of the house (0 = basement, 1 = first floor);
- County the house is located at;
- log_radon: Radon measurement from the house (log scale);
- log_uranium: Uranium level in the county (log scale).

Describe your proposed model, explaining each of the components, and provide estimates of the parameters following the Bayesian paradigm. Discuss your prior specification and point out potential limitations your model might have.

Question 2

This is exercise 12H1. from the book *Statistical Rethinking - A Bayesian Course with Examples in R and Stan* by Richard McElreath, 2nd edition. You need to install the `rethinking` package in `R` to access the data.

In 1980, a typical Bengali woman could have 5 or more children in her lifetime. By the year 200, 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 three of them for this practice problem:

- `district_id`: ID number of administrative district each woman resided in

- `use.contraception`: An indicator (0/1) of whether the woman was using contraception
- `urban`: An indicator (0/1) of whether the woman lived in a city, as opposed to living in a rural area.

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:

```
> sort(unique(d$district))
```

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))
```

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

Now, focus on predicting `use.contraception`, clustered by `district_id`. Do not include `urban` just yet. Fit both (1) a traditional fixed-effects model that uses dummy variables 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? In particular, can you explain the most extreme cases of disagreement, both why they happen where they do and why the models reach different inferences?