STA365 Homework 3

Daniel Simpson

2020-03-05

Please submit your answers as a **single pdf document** via Quercus. The documents should be prepared in an RMarkdown document. It will be ~~Due Tuesday 10 March, 2020 at 12pm (Midday)~~ **Due Tuesday 17 March at 12pm**. Late submissions will be heavily penalized.

This homework counts towards 5% of the final grade for the course. 1% will be awarded for each complete question. 1% will be awarded for the correct presentation of the figures (each figure should be captioned with informative captions and the axes and legends should be clearly labelled). 1% will be awarded for general presentation.

The radon in the basement

Radon is a carcinogen—a naturally occurring radioactive gas whose decay products are also radioactive—known to cause lung cancer in high concentrations and estimated to cause several thousand lung cancer deaths per year in the United States. The distribution of radon levels in U.S. homes varies greatly, with some houses having dangerously high concentrations. In order to identify the areas with high radon exposures, the Environmental Protection Agency coordinated radon measurements in a random sample of more than 80,000 houses throughout the country.

In performing the analysis, there is an important predictor—the floor on which the measurement was taken, either basement or first floor; radon comes from underground and can enter more easily when a house is built into the ground. There is also an important county-level predictor—a measurement of soil uranium that was only available at the county level.

To simplify the problem somewhat, our goal is to estimate the distribution of radon levels in each of the 85 counties in Minnesota. We expect the baseline level of radon radiation will vary from county to county, which will effect the intercept (ie the model will have a varying intercept). However we do not expect the effedct of floor will change from county to county.

Using this information, we get a model of the form

$$\begin{aligned} y_{ij} &\sim N(\alpha_j + \beta x_i, \sigma^2) \\ \alpha_j &\sim N(\gamma_0 + \gamma_1 u_j, \tau^2) \\ \beta, \gamma_0, \gamma_1 &\sim N(0, ?) \\ \sigma, \tau &\sim N_+(0, ?) \end{aligned}$$

where y_{ij} is the logarithm of the radon measurement in house i and county j, x is the floor of the measurement (that is, 0 for basement and 1 for first floor), and u is the logarithm of the uranium measurement at the county level.

Data for a subset of the data containing 919 observations in Minnesota are available in the `radon` data set in the `rstanarm` package and can be accessed by running the following code

```
# install.packages("rstanarm") ## If needed
library(rstanarm)
data(radon)
# You will need county as an integer to pass it to Stan
radon$county_int = as.integer(radon$county)
head(radon)
```

Questions

1. The "?" in the prior specification indicates values that need to be chosen (these values do not need to be the same). Choose sensible values of these variances and demonstrate that they are sensible by simulating from the prior predictive distribution. It may be useful to know that a value of 300 Becquerels per cubic meter (the units of `exp(radon$log_radon)`) would be consider extremely high.
2. Write a Stan program to fit a multilevel model with these priors chosen in part 1. When fitting the radon data it should run without any divergence, Rhat, or Effective Sample Size warnings. **Note:** You may need to do some data wrangling here!
3. Choose 5 counties and compare the estimated radon levels on floors 0 and 1 from the multilevel model in part 1 with a no-pooling regression estimate and comment on the difference.