HW 05: Geospatial Modeling

Mapping Anemia Across the Democratic Republic of Congo

> **!** Due date
>
> This assignment is due on **Tuesday, April 8 at 11:45am.** To be considered on time, the following must be done by the due date:
>
> - Final `.qmd` and `.pdf` files pushed to your GitHub repo
> - Final `.pdf` file submitted on Gradescope

Getting started

- Go to the biostat725-sp25 organization on GitHub. Click on the repo with the prefix **hw-05**. It contains the starter documents you need to complete the homework.
- Clone the repo and start a new project in RStudio. See the AE 01 instructions for details on cloning a repo and starting a new project in R.

Packages

The following packages are used in this assignment:

```
library(tidyverse)
library(rstan)
library(bayesplot)
library(knitr)
library(loo)
# new packages are below
library(sf)            # functions to work with spatial data
library(rnaturalearth) # maps of DRC
```

```
library(geodist)       # conversion from geodisic to miles

# load other packages as needed
```

Introduction

Data

We will look at a sample of women aged 15-49 sampled from the 2013-14 Democratic Republic of Congo (DRC) Demographic and Health Survey. There are ~8600 women who are nested in ~500 survey clusters. The variables in the dataset are as follows.

- `loc_id`: location id (i.e. survey cluster).
- `hemoglobin`: hemoglobin level (g/dL).
- `anemia`: anemia classifications.
- `age`: age in years.
- `urban`: urban vs. rural.
- `LATNUM`: latitude.
- `LONGNUM`: longitude.
- `mean_hemoglobin`: average hemoglobin at each community (g/dL).
- `community_size`: number of participants at each community.
- `mean_age`: average age of participants at each community (years).

The data set is the same one that was used in AE-07 and available in your HW repos.

```
drc <- readRDS("drc.rds")
```

Modeling

In this homework, we will focus on the Sud-Kivu state within the DRC. Researchers are interested in mapping risk of anemia across Sud-Kivu. Anemia is a condition in which the body lacks enough healthy red blood cells to carry adequate oxygen to tissues. It is commonly associated with **low levels of hemoglobin**, the oxygen-carrying protein in red blood cells. Anemia can result from malnutrition, especially deficiencies in iron, folate, or vitamin B12, and is linked to a range of negative outcomes including fatigue, impaired cognitive development, weakened immunity, and poor pregnancy outcomes. In children and women, anemia is often a marker of underlying nutritional and health disparities.

To study anemia, researchers would like to fit a logistic regression model using a Bayesian hierarchical model with spatially varying intercepts. Define $Y_{ij} \in \{0, 1\}$ as the indicator of anemia at location i ($i = 1, \dots, n$) for participant j ($j = 1, \dots, n_i$).

```
drc <- drc %>%
  mutate(anemic = if_else(anemia == "not anemic", 0, 1))
```

We will fit the following model:

$$
\begin{aligned}
Y_j(\mathbf{u}_i)|\alpha, \beta, \theta(\mathbf{u}_i) &\overset{ind}{\sim} \text{Bernoulli}(\pi_j(\mathbf{u}_i)) \\
\text{logit}(\pi_j(\mathbf{u}_i)) &= \alpha + \mathbf{x}_j(\mathbf{u}_i)\beta + \theta(\mathbf{u}_i) \\
\theta|\tau, \rho &\sim N(\mathbf{0}_n, \mathbf{C}) \\
\alpha^* &\sim N(0, 4^2) \\
\beta_j|\sigma_\beta &\sim N(0, \sigma_\beta^2), \quad j = 1, \dots, p \\
\tau &\sim \text{Half-Normal}(0, 4^2) \\
\rho &\sim \text{Inv-Gamma}(5, 5) \\
\sigma_\beta &\sim \text{Half-Normal}(0, 2^2),
\end{aligned}
$$

where $\theta = (\theta(\mathbf{u}_1), \dots, \theta(\mathbf{u}_n))^\top$, $N = \sum_{i=1}^n n_i = 490$, $n = 29$, $\mathbf{x}_j(\mathbf{u}_i) = (\text{age}_{ij}/10, \text{urban}_i)$, and $\mathbf{C}$ a covariance matrix produced using the Matérn 3/2 covariance function with scale parameter τ and length scale ρ. The parameter α^* is the intercept after centering, such that $\text{logit}(\pi_j(\mathbf{u}_i)) = \alpha^* + \mathbf{x}_j^*(\mathbf{u}_i)\beta + \theta(\mathbf{u}_i)$, where $\mathbf{x}_j^*(\mathbf{u}_i) = \mathbf{x}_j(\mathbf{u}_i) - \bar{\mathbf{x}}_j(\mathbf{u}_i)$ and $\bar{\mathbf{x}}_j(\mathbf{u}_i) = \frac{1}{N}\sum_{i=1}^n \sum_{j=1}^{n_i} \mathbf{x}_j(\mathbf{u}_i)$.

Exercises

Exercise 1

Based on AE-07, process the `drc` data to create an `sf` data object that can be used for spatial analyses. Create a map that visualizes the proportion of participants with anemia at each

community in Sud-Kivu.

Exercise 2

Create a 25×25 grid of points to be used for prediction, and only keep those points within Sud-Kivu. Visualize the grid.

Exercise 3

Researchers are interested in understanding the **a priori** correlation between the anemia probability between two communities across Sud-Kivu. Visualize the correlation as a function of distance between two locations given in miles, evaluated at the prior mean for ρ. What is the a priori correlation between two locations that are 25 miles apart?

Exercise 4

Fit the model detailed above. Present evidence of MCMC convergence.

Exercise 5

Researchers would like to understand the **a posteriori** correlation between the anemia probability between two communities across Sud-Kivu. Repeat **Exercise 3** using the posterior mean of ρ. Visualize the correlations from the prior and posterioron the same figure. How does the posterior correlation compare to the prior?

Exercise 6

Present posterior means and intervals for the odds ratios for age and urbanality. Are any of these predictors associated with anemia?

Exercise 7

Map the posterior mean probability of anemia across your Sud-Kivu, $\pi(\mathbf{u}_i)$. To make this map, we must specify $x(\mathbf{u}_i) = (2.8, 0)$, such that the age is 28 years (average age in Sud-Kivu) and we assume locations are rural (most common community location). Describe any spatial patterns that arise.

Exercise 8

Create a map of the posterior standard deviation that corresponds to your map in **Exercise 5**. Describe any spatial patterns that arise.

Exercise 9

Researchers are worried about any locations where the posterior probability is greater than 0.4, $p_i = P(\pi(\mathbf{u}_i) > 0.4|\mathbf{Y})$. In particular, any locations with p_i greater than 0.5 ($h_i = 1(p_i > 0.5)$) indicates a hot spot region where public health individuals would like to intervene. Visualize both the p_i and h_i and make a statement about where intervention is warranted.

Submission

You will submit the PDF documents for homeworks, and exams in to Gradescope as part of your final submission.

> **Warning**
>
> Before you wrap up the assignment, make sure all documents are updated on your GitHub repo. We will be checking these to make sure you have been practicing how to commit and push changes.
> Remember – you must turn in a PDF file to the Gradescope page before the submission deadline for full credit.

To submit your assignment:

- Access Gradescope through the menu on the BIOSTAT 725 Canvas site.
- Click on the assignment, and you'll be prompted to submit it.
- Mark the pages associated with each exercise. All of the pages of your homework should be associated with at least one question (i.e., should be "checked").

Grading

Component	Points
Ex 1	5
Ex 2	10
Ex 3	5

Component	Points
Ex 4	5
Ex 5	5
Ex 6	5
Ex 7	5
Ex 8	5
Ex 9	5