SCI 2025: Homework 2

Setup

In this homework, we will work with data from:

Nettle, D. (1998). Explaining global patterns of language diversity. *Journal of Anthropological Archaeology*, 17:354–74.

First, load the data into your R session.

```
library(rethinking)
data(nettle)
head(nettle)
```

```
     country num.lang    area  k.pop num.stations mean.growing.season
1    Algeria       18 2381741  25660          102                6.60
2     Angola       42 1246700  10303           50                6.22
3  Australia      234 7713364  17336          134                6.00
4 Bangladesh       37  143998 118745           20                7.40
5      Benin       52  112622   4889            7                7.14
6    Bolivia       38 1098581   7612           48                6.92
  sd.growing.season
1              2.29
2              1.87
3              4.17
```

```
4               0.73
5               0.99
6               2.50
```

The meaning of each column in the dataset is given below:

(1) country: Name of the country
(2) num.lang: Number of recognized languages spoken
(3) area: Area in square kilometers
(4) k.pop: Population, in thousands
(5) num.stations: Number of weather stations that provided data for the next two columns
(6) mean.growing.season: Average length of growing season,in months
(7) sd.growing.season: Standard deviation of length of growing season,in months

You should use quadratic approximation via `rethinking::quap()` for all model fitting.

Question 1

Write down a mathematical model that describes a linear regression of the number of languages spoken (`num.lang`) as a function of the population of the country (`k.pop`). Use similar notation to the textbook chapter. Be sure to include prior definitions for all parameters. You may apply any transformations to the data that you think are appropriate.

Question 2

Implement the model you wrote down in Question 1 using `rethinking::quap()`. Print the model summary.

Question 3

Perform a posterior predictive check on the model you fit in Question 2. You should plot the posterior function relating the number of languages spoken to the population of the country. Represent uncertainty either by drawing lines from the posterior or by plotting a credible/highest posterior density interval. Be sure to also plot the raw data.

Question 4

Visually compare the prior and posterior distributions of the *parameters* from the model you fit in Question 2.

Question 5

Using insights from Questions 3-4, try to improve upon the model you fit in Question 2. Justify your changes, fit the new model, and perform a new posterior predictive check.