ESS 575: Bayesian Regression Lab

Team England

12 October, 2022

Team England:

- Caroline Blommel
- Carolyn Coyle
- Bryn Crosby
- George Woolsey

cblommel@mail.colostate.edu, carolynm@mail.colostate.edu, brcrosby@rams.colostate.edu, george.woolsey@colostate.edu

Setup

In this lab you will practice building regression models that are appropriate for a given data set.

- Coexistance data
- Abundance data

Load the data

```
coexist_pth <- "https://nthobbs50.github.io/ESS575/content/labs/coexist.csv"
coverage_pth <- "https://nthobbs50.github.io/ESS575/content/labs/hesp_coverage.csv"

coexist <- read.csv(coexist_pth)
coverage <- read.csv(coverage_pth)
```

Problem A.

Hein et. al (2012) investigated how environmental conditions influence coexistence between Arctic char (*Salmo trutta*) and pike (*Esox lucius*) in Swedish lakes. Pike were introduced to 151 lakes containing brown trout. Coexistence of the two species was recorded, as $yi = 1$ if both species were found in lake i and 0 otherwise. For each lake, five environmental conditions deemed relevant to coexistence patterns were observed:

- elevation
- upstream catchment area

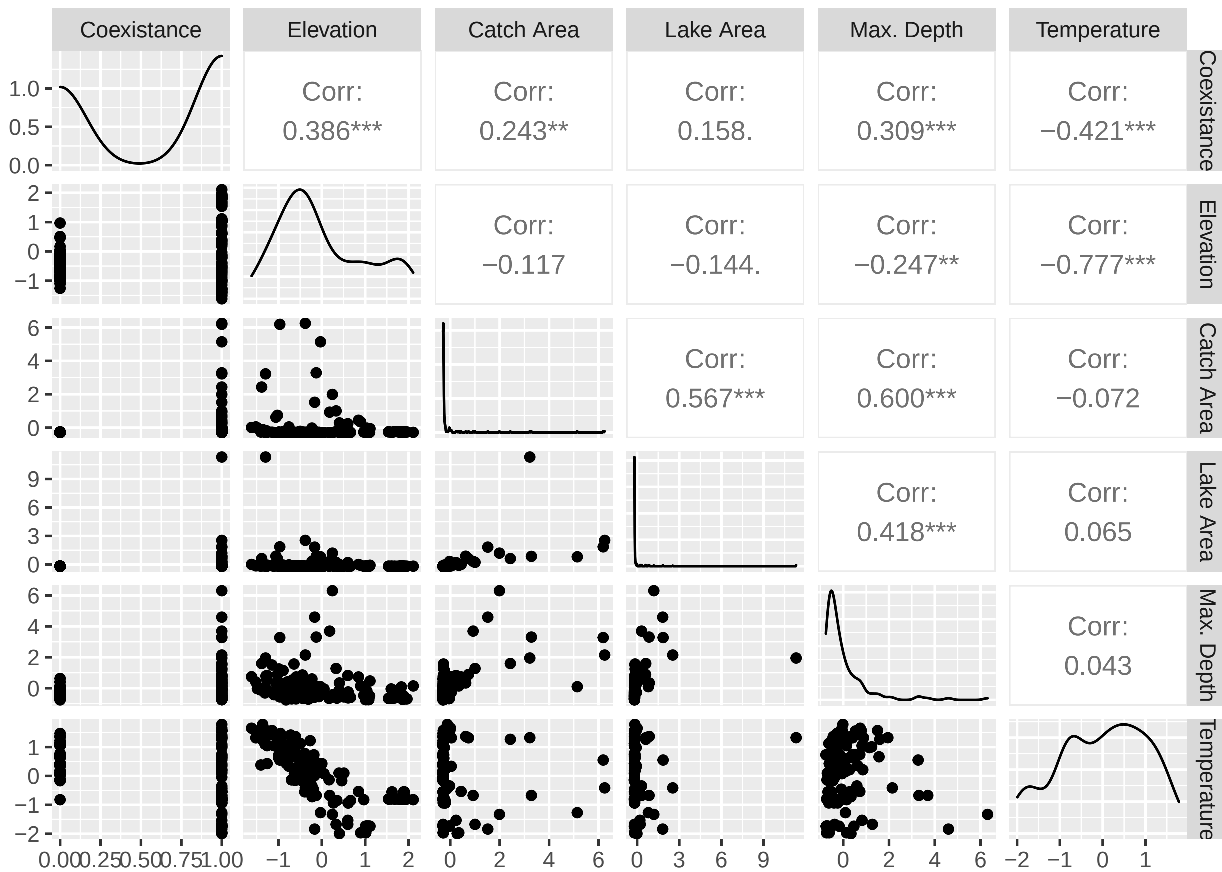
- maximum area
- maximum depth
- mean annual air temperature at outlet

The predictors have been standardized.

Question 1

Plot the data using `pairs` or `GGally::ggpairs`. What do you notice? How might this impact your modeling choices?

```
GGally::ggpairs(
  data = coexist
  , columns = c("coexist", "elev", "catcharea", "lakearea", "maxdepth", "temp1")
  , columnLabels = c("Coexistance", "Elevation", "Catch Area", "Lake Area", "Max. Depth", "Temperature")
)
```



This grid of plots given by `GGally` shows the empirical density (a.k.a. marginal distribution) of each variable on the diagonal, the scatterplot of points for pairs of variables on the lower triangle, and the Pearson correlation between variables in the upper right triangle.

We notice that the coexist outcome variable takes on two values (0 and 1). We also notice that there is one relatively large lake in the data set. A binary data model is appropriate given the support of the data.

Question 2

We seek to understand the relationship between the probability of coexistence and the five environmental covariates. Write out a reasonable data model.

$$y_i \sim \mathsf{Bernoulli}\big(p_i\big)$$

where:

$$p_i = g(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, x_i) = \text{inverse logit}\big(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_5 x_{5i}\big) = \frac{\exp\big(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_5 x_{5i}\big)}{1 + \exp\big(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_5 x_{5i}\big)}$$

Question 3

Assume we know very little about the impact of the environmental covariates on coexistence. Specify appropriate prior distributions for all unknown parameters. Explain your choice of priors.

Assuming vague priors on the intercept and slope we will set the variance $\sigma^2 = 2.7$ normally distributed with a mean of 0; e.g., $\beta_0 \sim \mathsf{normal}(0, 2.7),\ \beta_1 \sim \mathsf{normal}(0, 2.7), \cdots,\ \beta_1 \sim \mathsf{normal}(0, 2.7)$. Such that:

$$\big[\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5 \mid \mathbf{y}\big] \propto \prod_{i=1}^{n} \mathsf{Bernoulli}\big(y_i \mid g(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, x_i)\big) \times \mathsf{normal}\big(\beta_0 \mid 0, 2.7\big) \times \cdots \times \mathsf{normal}\big(\beta_5 \mid 0, 2.7\big)$$