HW 02: Bayesian linear regression

Physical activity and access to recreational facilities in Cook County, Chicago

> **!** Due date
>
> This assignment is due on **Thursday, February 12 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-sp26 organization on GitHub. Click on the repo with the prefix **hw-02**. 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)

# load other packages as needed
```

Introduction

It is estimated that only 3.0% of Americans engage in a fully healthy lifestyle, which entails refraining from smoking, eating five or more fruits and vegetables daily, maintaining a healthy weight, and participating in regular exercise (a component of physical activity). Lack of physical activity is a leading risk factor for chronic disease and having access to recreational facilities has been associated with an individual's level of physical activity and exercise. Researchers were interested in analyzing the association between access to recreational facilities and exercise, controlling for crime. It has been suggested that reductions of violence/crime and increased perceptions of neighborhood safety may contribute to higher population levels of physical activity.

Researchers performed a cross-sectional observational study of Cook County, Chicago, randomly recruiting 87 pregnant women and obtained an estimate of weekly exercise as measured by metabolic equivalent (MET) minutes per week. Measures of neighborhood recreational facilities and crime were obtained based on geographical kernel estimates. The data they collected can be found in the dataset, *exercise.csv*, which is available in the homework repo.

The following variables are in the dataset:

- `exercise`: exercise measured in metabolic equivalent (MET) minutes per week.
- `recreation`: number of recreational facilities within a one-mile radius of a participant's home.
- `crime`: total yearly average of all crimes in a one-mile buffer surrounding an individual's residence per 1,000 persons.
- `age`: age in years.
- `married`: marriage status (1 = married; 0 = single)
- `race`: race (0 = White ; 1 = African-American/black, 2 = Asian)

Exercise 1

Setup a multivariable linear regression to estimate the association between access to recreational facilities and exercise, making sure to allow for this relationship to change based on crime. Be sure to control for the following confounders: age, marital status, and race. Define the random variable Y_i as the exercise in MET minutes per week for women i and assume that $Y_i \overset{ind}{\sim} N(\mu_i, \sigma^2)$ for $i = 1, \ldots, n$ where

$$\begin{aligned}\mu_i &= \alpha + recreation_i\beta_1 + crime_i\beta_2 + (recreation_i \times crime_i)\beta_3 \\ &\quad + age_i\beta_4 + black_i\beta_5 + asian_i\beta_6 + married_i\beta_7 \\ &= \alpha + \mathbf{x}_i\beta.\end{aligned}$$

Fit this regression using a Bayesian framework in Stan to estimate (α, β, σ). For all model parameters, choose weakly-informative priors, $\alpha \sim N(0, 100)$, $\beta_j \sim N(0, 100)$ for $j = 1, \ldots, p$, and $\sigma \sim$ Half-Normal$(0, 100)$. Recall, that in this class, normal distributions are parameterized using variances (i.e., the standard deviations in these priors is 10).

Evaluate model convergence and present posterior predictive checks. Provide an argument for whether the model fits the data well.

Exercise 2

Refit the model in **Exercise 1**, this time using centered outcome and predictor variables, $Y_i^* \overset{ind}{\sim} N(\mu_i, \sigma^2)$, where $\mu_i^* = \alpha + \mathbf{x}_i^* \beta$. The centered data are defined as, $Y_i^* = Y_i - \overline{Y}$, where $\overline{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$ and $\mathbf{x}_i^* = \mathbf{x}_i - \bar{\mathbf{x}}$, where $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$. Use the same priors as before. Once again, evaluate model convergence and present posterior predictive checks. Provide an argument for whether the model fits the data well and make a comparison to the model from **Exercise 1**. Make an argument for why this model may have improved model fit.

Exercise 3

For the model from **Exercise 2**, present the posterior means, standard deviations, and 95% credible intervals for all model parameters. What is interpretation of the slope main effect corresponding to `recreation` in your model? Within the context of the association between access to recreational facilities and exercise, is this main effect parameter useful to interpret?

Exercise 4

What is the association between access to recreational facilities and exercise, for a pregnant women living in an area with 5 annual crimes/1,000 people? What about for 15 annual crimes/1,000 people? Provide posterior mean and standard deviations for both quantities.

Exercise 5

Interpret the posterior mean and standard deviations from **Exercise 4** and compare and contrast them. What do these posterior slopes say about the impact of crime on the relationship between access to recreational facilities and exercise.

Exercise 6

Researchers are interested in the level of crime where the association between recreational facilities and exercise disappears. Present the posterior median and interquartile range (i.e., 25% and 75% percentiles) for this quantity.

Exercise 7

Compute the posterior predictive distribution for a patient with 10 recreational facilities within a one-mile radius, 5 crimes within a one-mile buffer per 1,000 people, is 40 years old, married, and white race. Report the posterior mean and 95% credible intervals.

Exercise 8

Researchers are interested in comparing their original model (i.e., the one from **Exercise 2**) with a model that does not contain an interaction term between recreational facility access and crime. Fit the model without the interaction term and perform a model comparison between the two models using an information criteria. Which model would you suggest as more scientifically plausible?

Exercise 9

Perform a sensitivity analysis to the choice of prior for (α, β, σ). Make sure to change the family of priors for each parameter. Are your results robust to the choice of prior?

You're done and ready to submit your work! render, commit, and push all remaining changes. You can use the commit message "Done with Homework 2!", and make sure you have pushed all the files to GitHub (your Git pane in RStudio should be empty) and that all documents are updated in your repo on GitHub. The PDF document you submit to Gradescope should be identical to the one in your GitHub repo.

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	8
Ex 2	8
Ex 3	4
Ex 4	4
Ex 5	3
Ex 6	5
Ex 7	6
Ex 8	7
Ex 9	5