Homework 1

Problem 0

This "problem" focuses on structure of your submission, especially the use git and GitHub for reproducibility, R Projects to organize your work, R Markdown to write reproducible reports, relative paths to load data from local files, and reasonable naming structures for your files.

To that end:

- create a public GitHub repo + local R Project; I suggest naming this repo / directory bios731_hw1_YourLastName (e.g. bios731_hw1_wrobel for Julia)
- Submit your whole project folder to GitHub
- Submit a PDF knitted from Rmd to Canvas. Your solutions to the problem here should be implemented in your .Rmd file, and your git commit history should reflect the process you used to solve these Problems.

Problem 1

Simulation study: our goal in this homework will be to plan a well-organized simulation study for multiple linear regression and bootstrapped confidence intervals.

Below is a multiple linear regression model, where we are interested in primarily treatment effect.

$$Y_i = \beta_0 + \beta_{treatment} X_{i1} + \mathbf{Z_i}^T \boldsymbol{\gamma} + \epsilon_i$$

Notation is defined below:

- Y_i: continuous outcome
- X_{i1}: treatment group indicator; $X_{i1} = 1$ for treated
- $\mathbf{Z_i}$: vector of potential confounders
- $\beta_{treatment}$: average treatment effect, adjusting for $\mathbf{Z_i}$
- $\boldsymbol{\gamma}$: vector of regression coefficient values for confounders
- ϵ_i: errors, we will vary how these are defined

In our simulation, we want to

- Estimate $\beta_{treatment}$ and $se(\hat{\beta}_{treatment})$
  - Evaluate $\beta_{treatment}$ through bias and coverage
  - We will use 3 methods to compute $se(\hat{\beta}_{treatment})$ and coverage:
    1. Wald confidence intervals (the standard approach)
    2. Nonparametric bootstrap percentile intervals
    3. Nonparametric bootstrap t intervals
  - Evaluate computation times for each method to compute a confidence interval
- Evaluate these properties at:

  - Sample size $n \in \{10, 50, 500\}$
  - True values $\beta_{treatment} \in \{0, 0.5, 2\}$
  - True ϵ_i normally distributed with $\epsilon_i \sim N(0, 2)$
  - True ϵ_i coming from a right skewed distribution
    * **Hint**: try $\epsilon_i \sim logNormal(0, \log(2))$
- Assume that there are no confounders ($\boldsymbol{\gamma} = 0$)
- Use a full factorial design

Problem 1.1 ADEMP Structure

Answer the following questions:

- How many simulation scenarios will you be running?

There are $3 * 3 * 2 = 18$ simulation scenarios that I will be running using a full factorial design.

- What are the estimand(s)?

The estimands are $\beta_{treatment}$ and $se(\hat{\beta}_{treatment})$.

- What method(s) are being evaluated/compared?

We are comparing 3 methods to compute the standard error of the estimated regression coefficients, $\hat{\beta}_{treatment}$: Wald CIs, Nonparametricbootstrap percentile intervals and nonparametric bootstrap t intervals.

- What are the performance measure(s)?

The performance measures are bias, coverage, and computation time.

Problem 1.2 nSim

Based on desired coverage of 95% with Monte Carlo error of no more than 1%, how many simulations (n_{sim}) should we perform for each simulation scenario? Implement this number of simulations throughout your simulation study.

```
coverage <- 0.95
MCE <- 0.01
nsim <- (coverage*(1-coverage))/(MCE^2)
```

We should perform 475 simulations for each simulation scenario.

```
n = c(10, 50, 100)
beta_true = c(0, 0.5, 2)
sigma2_true = c("normal", "skewed")

params = expand.grid(n = n,
                     n_sim = nsim,
                     beta_true = beta_true,
                     sigma2_true = sigma2_true)
```

Problem 1.3 Implementation

We will execute this full simulation study. For full credit, make sure to implement the following:

- Well structured scripts and subfolders following guidance from `project_organization` lecture
- Use relative file paths to access intermediate scripts and data objects
- Use readable code practices
- Parallelize your simulation scenarios
- Save results from each simulation scenario in an intermediate `.Rda` or `.rds` dataset in a `data` subfolder
  - Ignore these data files in your `.gitignore` file so when pushing and committing to GitHub they don't get pushed to remote
- Make sure your folder contains a Readme explaining the workflow of your simulation study
  - should include how files are executed and in what order
- Ensure reproducibility! I should be able to clone your GitHub repo, open your `.Rproj` file, and run your simulation study

Problem 1.4 Results summary

Create a plot or table to summarize simulation results across scenarios and methods for each of the following.

- Bias of $\hat{\beta}$
- Coverage of $\hat{\beta}$
- Distribution of $se(\hat{\beta})$
- Computation time across methods

If creating a plot, I encourage faceting. Include informative captions for each plot and/or table.

Problem 1.5 Discussion

Interpret the results summarized in Problem 1.4. First, write a **paragraph** summarizing the main findings of your simulation study. Then, answer the specific questions below.

- How do the different methods for constructing confidence intervals compare in terms of computation time?
- Which method(s) for constructing confidence intervals provide the best coverage when $\epsilon_i \sim N(0, 2)$?
- Which method(s) for constructing confidence intervals provide the best coverage when $\epsilon_i \sim logNormal(0, \log(2))$?