Homework 1

Context

This assignment reinforces ideas in Module 1: Reproducible computing in R. We focus specifically on implementing a large scale simulation study, but the assignment will also include components involving bootstrap and parallelization, Git/GitHub, and project organization.

Due date and submission

Please submit (via Canvas) a PDF knitted from .Rmd. Your PDF should include the web address of the GitHub repo containing your work for this assignment; git commits after the due date will cause the assignment to be considered late.

R Markdown documents included as part of your solutions must not install packages, and should only load the packages necessary for your submission to knit.

Points

Problem	Points
Problem 0	20
Problem 1.1	10
Problem 1.2	5
Problem 1.3	20
Problem 1.4	30
Problem 1.5	15

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? **ANSWER:** 3 x 3 x 2 = 18 simulation scenarios.
- What are the estimand(s) **ANSWER:** The average treatment effect, $\beta_{treatment}$.
- What method(s) are being evaluated/compared? **ANSWER:** Wald confidence intervals, Nonparametric bootstrap percentile intervals, and Nonparametric bootstrap t intervals.
- What are the performance measure(s)? **ANSWER:** 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.

```
0.95 * (1 - 0.95) / (0.01^2)
```

```
## [1] 475
```

```
# We need 475 simulations per scenario.
```

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

```
# Run this code to run the simulations
source(here::here("11_simulations", "run_simulations.R"))
```

Note: simulations were taking a long time, so I used 50 simulations for each scenario instead of 475. I also used too 50 bootstrap samples and 25 bootstrap samples for the inner bootstrap for bootstrap t intervals. If I had more compute time, I'd take 1,000 bootstrap samples and 500 samples for the inner bootstrap.

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.

Table 2: Summary of Simulation Results

	Bias	Bias 95% CI	St. Err.	St. Err. 95% CI	Cover Wald	Cover Pct	Cover BSt
Sc 1	-0.14000	(-2.56, 1.96)	0.811	(0.484, 1.25)	0.78	0.76	0.90
Sc 2	0.02310	(-0.782, 0.725)	0.395	(0.305, 0.616)	0.96	0.88	0.96
Sc 3	-0.02910	(-0.493, 0.573)	0.274	(0.221, 0.345)	0.94	0.88	0.92
Sc 4	-0.01170	(-1.84, 1.92)	0.802	(0.424, 1.35)	0.84	0.82	0.90
Sc 5	-0.09060	(-0.588, 0.557)	0.397	(0.299, 0.499)	0.98	0.94	0.94
Sc 6	-0.03860	(-0.476, 0.388)	0.275	(0.226, 0.333)	1.00	0.98	0.98
Sc 7	-0.00834	(-1.46, 1.49)	0.894	(0.463, 1.37)	0.88	0.82	0.96
Sc 8	0.01050	(-0.788, 0.78)	0.390	(0.304, 0.481)	0.92	0.90	0.90
Sc 9	0.03240	(-0.534, 0.472)	0.283	(0.21, 0.336)	0.96	0.88	0.92
Sc 10	-0.86400	(-13.5, 6.58)	3.490	(0.368, 13.5)	0.92	0.76	0.84
Sc 11	-0.11600	(-3.71, 2.21)	1.430	(0.527, 3.04)	0.98	0.94	0.94
Sc 12	0.03900	(-3.45, 2.83)	1.160	(0.525, 2.63)	0.96	0.86	0.88
Sc 13	-0.00172	(-6.39, 5.22)	2.400	(0.518, 7.46)	0.94	0.88	0.92
Sc 14	0.23000	(-3.39, 4.46)	1.580	(0.451, 6.01)	1.00	0.90	0.90
Sc 15	-0.17300	(-2.78, 2.05)	1.120	(0.473, 2.35)	0.94	0.86	0.80
Sc 16	-1.10000	(-10, 7.6)	3.390	(0.507, 9.7)	0.94	0.84	0.88
Sc 17	-0.20300	(-3.09, 2.04)	1.370	(0.549, 3.58)	0.96	0.92	0.86
Sc 18	0.02460	(-1.85, 2.67)	1.290	(0.551, 3.28)	1.00	0.94	0.86

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?

  I did not have the time to run the methods using separate bootstraps to evaluate computation time. I computed the confidence intervals for all three methods using the same bootstrap samples.

- Which method(s) for constructing confidence intervals provide the best coverage when $\epsilon_i \sim N(0, 2)$?

  The Wald method.

- Which method(s) for constructing confidence intervals provide the best coverage when $\epsilon_i \sim logNormal(0, \log(2))$?

  The Wald method.

Main Findings

The simulation study evaluated the bias and standard error of the average treatment effect by assuming three different values of the treatment effect, three different sample sizes, and two different error distributions. We compared the performance of three methods for constructing confidence intervals. The results suggest that the Wald method frequently provides the highest coverage, while the percentile method tends to be the most conservative. Additionally, misspecifying the error distribution (secnarios 10 through 18) does not introduce bias in the estimation of the treatment effect, but it does affect the standard error.