Homework 2: A simulation study investigating the bootstrap

BIOS 731 – Advanced Statistical Computing

Context and learning objectives

This assignment reinforces ideas in Module 2: **Simulations and Resampling Methods**. We focus specifically on implementing a **large-scale simulation study**, and the assignment also includes components involving the bootstrap, parallelization, Git/GitHub, and project organization.

Due Date and Submission

Submit (via Canvas) a PDF knitted from a `.Rmd` or `.qmd` file. Your PDF should include the web address of the GitHub repository containing your work for this assignment. **Commits after the due date will cause the assignment to be considered late.**

Point distribution

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 the structure of your submission, especially the **use of git and GitHub** for reproducibility, **R Projects** to organize your work, **Quarto/R Markdown** to write reproducible reports, **relative paths** to load local files, and reasonable naming conventions for your files.

To that end:

- Create a public GitHub repository and a local R Project. I suggest naming the repo/directory `bios731_hw2_YourLastName` (e.g., `bios731_hw2_wrobel`).
- Push your **entire project folder** to GitHub.
- Submit a PDF knitted from your `.Rmd/.qmd` file to Canvas.
  - Your solutions should be implemented in your `.Rmd/.qmd` file.
  - Your git commit history should reflect your workflow (i.e., multiple meaningful commits; avoid a single "final" commit with all work).

Problem 1

Simulation study. Your goal in this homework is to plan and execute a well-organized simulation study for multiple linear regression and confidence intervals constructed via both Wald and bootstrap methods.

Model

Consider the multiple linear regression model

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

where we are primarily interested in the treatment effect $\beta_{treatment}$. It is fine to simulate data with no confounders, i.e. $\gamma = 0$.

Notation

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 $\mathbf{Z_i}$
- γ: vector of regression coefficient values for confounders

- ϵ_i: errors, we will vary how these are defined

Simulation goals

In our simulation, we want to

- Estimate $\beta_{treatment}$ and $se(\hat{\beta}_{treatment})$
  - Evaluate $\beta_{treatment}$ through bias and coverage

You will compare **three** methods for constructing a 95% confidence interval for $\hat{\beta}_{treatment}$:

1. Wald confidence intervals (standard model-based approach)
2. Nonparametric bootstrap **percentile** intervals
3. Nonparametric bootstrap t intervals

You will also evaluate **computation time** for each method.

Simulation design (full factorial)

Evaluate performance across the following factors:

- Sample size: $n \in \{10, 50, 500\}$
- True treatment effect: $\beta_{treatment} \in \{0, 0.5, 2\}$
- Error distribution:
  - Normal errors: $\epsilon_i \sim N(0, 2)$
  - Heavy-tailed errors: $\epsilon_i \sim t_\nu$ with $\nu = 3$, scaled to have variance 2

Implementation hint (heavy tails). If $u \sim t_{nu}$ with $\nu > 2$, then $\text{Var}(u) = \nu/(\nu - 2)$. To match the normal-error condition variance, set

$$\epsilon_i = u \cdot \sqrt{2\,\frac{\nu - 2}{\nu}}.$$

Problem 1.1 ADEMP Structure

Answer the following questions. Use the ADEMP framework explicitly:

- **A (Aim):** What is the goal of the simulation study?
- **D (Data-generating mechanism):** What model and distributions generate the data? What factors vary across scenarios?
- **E (Estimand):** What quantity(ies) are you trying to learn about?
- **M (Methods):** What methods are being evaluated/compared?
- **P (Performance measures):** What metrics summarize performance?

Also answer:

- How many simulation scenarios will you be running (i.e., how many unique combinations in the full factorial design)?

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 you perform for **each** simulation scenario? Implement this value of n_{sim} throughout your simulation study.

Problem 1.3 Implementation

For bootstrap t, goal is to estimate a t distribution given by

$$t^* = \frac{\hat{\theta}^* - \hat{\theta}}{s_{\hat{\theta}^*}}$$

where

- $\hat{\theta}^*$: estimated parameter value from each bootstrap iteration
- $\hat{\theta}$: parameter estimate from the original sample
- $s_{\hat{\theta}^*}$ standard error estimate from a given bootstrap sample; requires a second level of bootstrapping to construct.

Parameter choices

- Use **B = 500** outer bootstrap resamples for the percentile and bootstrap-t intervals.
- For bootstrap-t, use **B_inner = 100** inner bootstrap resamples.
- Construct **95%** confidence intervals for all methods.

(If you choose different values, justify your choice and discuss the computation/accuracy trade-off.)

Computing + reproducibility requirements

Execute the full simulation study. For full credit, implement the following:

- Well-structured scripts and subfolders following guidance from the `project_organization` lecture
- Use relative file paths to access intermediate scripts and data objects
- Use readable code practices (clear function boundaries, meaningful names, minimal duplication)
- **Parallelize across simulation scenarios**
- Save results from each simulation scenario to an intermediate `.Rds` or `.Rda` file in a `data/` subfolder
  - Add these files to `.gitignore` so they are not pushed to GitHub
- Include a `README.md` explaining your workflow
  - Include what files to run, in what order, and how outputs are produced
- Ensure end-to-end reproducibility:
  - I should be able to clone your GitHub repo, open your `.Rproj`, and run the simulation study to regenerate results

Problem 1.4 Results summary

Create a plot or table summarizing simulation results across scenarios and methods for each of the following:

- Bias of $\hat{\beta}$
- Coverage of the **95% CI** for $\hat{\beta}$
- Distribution of $se(\hat{\beta})$
- Computation time across methods

Presentation guidance

- If creating plots, I encourage faceting by at least one design factor (e.g., n, error distribution, or true treatment value).
- Include informative captions for each plot/table.
- For coverage plots, consider adding a reference line at 0.95.

Problem 1.5 Discussion

Interpret the results summarized in Problem 1.4.

1. Write **one paragraph** summarizing the main findings of your simulation study.
2. Then answer the questions below:

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

Finally, briefly comment on any notable interactions (e.g., how performance changes with n or error type) and any practical recommendations you would make based on your results.