Homework 3

Context

This assignment reinforces ideas in Module 3: Cluster computing.

Due date and submission

Please submit (via Canvas) a PDF containing a link to 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.

Points

Problem	Points
Problem 0	20
Problem 1	80

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
- Submit your whole project folder to GitHub
- Submit a PDF knitted from Rmd to Canvas. Your solutions to the problems here should be implemented in your .Rmd file, and your git commit history should reflect the process you used to solve these Problems.

The Github address: https://github.com/ytliu36/bios731_hw3_liu.git

Problem 1

Continuation of Homework 1. Here, we will re-run part of the simulation study from Homework 1 with some minor changes, on the cluster. Cluster computing space is limited so we will not run too many jobs or simulations.

Problem 1 setup

The simulation study is specified below:

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 $\beta_{treatment}$
  - Evaluate $\beta_{treatment}$ through bias, coverage, type 1 error, and power
  - We will use 2 methods to compute $se(\hat{\beta}_{treatment})$ and coverage:
    1. Wald confidence intervals (the standard approach)
    2. Nonparametric bootstrap percentile intervals
  - Evaluate computation times for each method to compute a confidence interval
- Evaluate these properties at:
  - Sample size $n = \{20\}$
  - True values $\beta_{treatment} \in \{0, 0.5\}$
  - True ϵ_i normally distributed with $\epsilon_i \sim N(0, 2)$
  - True ϵ_i coming from a highly right skewed distribution
    - Generate data from a Gamma distribution with `shape = 1` and `rate = 2`.
- Assume that there are no confounders ($\boldsymbol{\gamma} = 0$)
- Use a full factorial design
- Use same `nsim` as previous assignment.

Problem 1 tasks

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

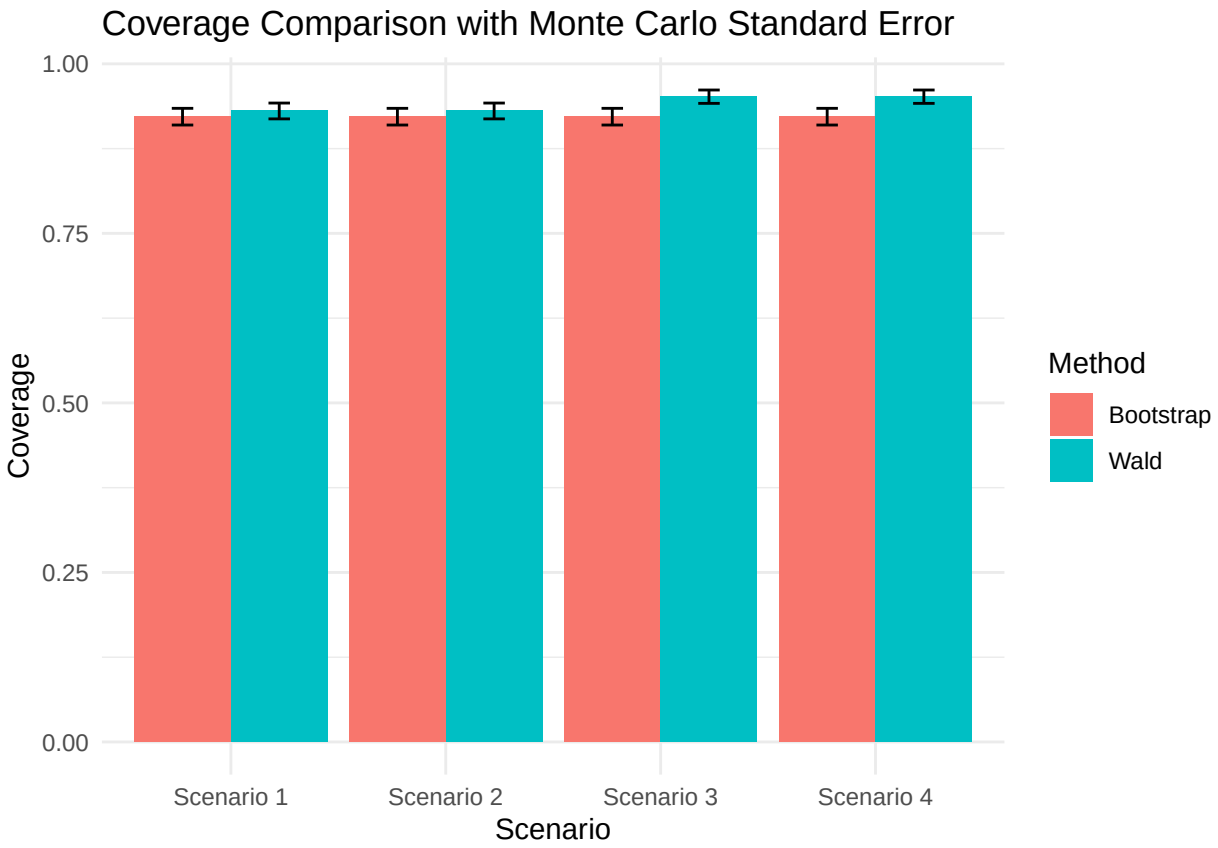
Workflow: * Use structured scripts and subfolders following guidance from the cluster computing project organization lecture * Instead of parallelizing your simulation scenarios (as in HW1), each simulation scenario should be assigned a different JOBID on the cluster.

Presenting results:

Create plots with *Monte Carlo standard error bars* to summarize the following:

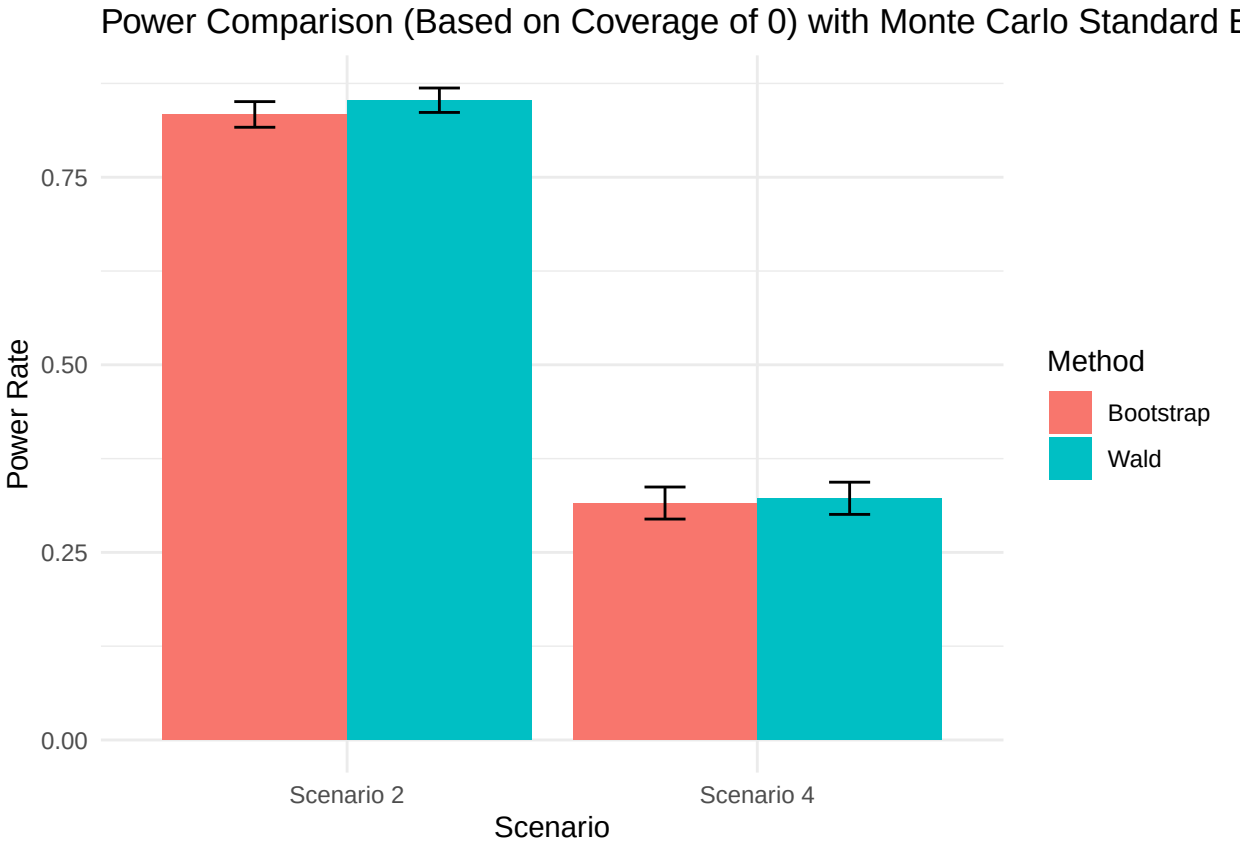
- Bias of $\hat{\beta}$

- Coverage of $\hat{\beta}$



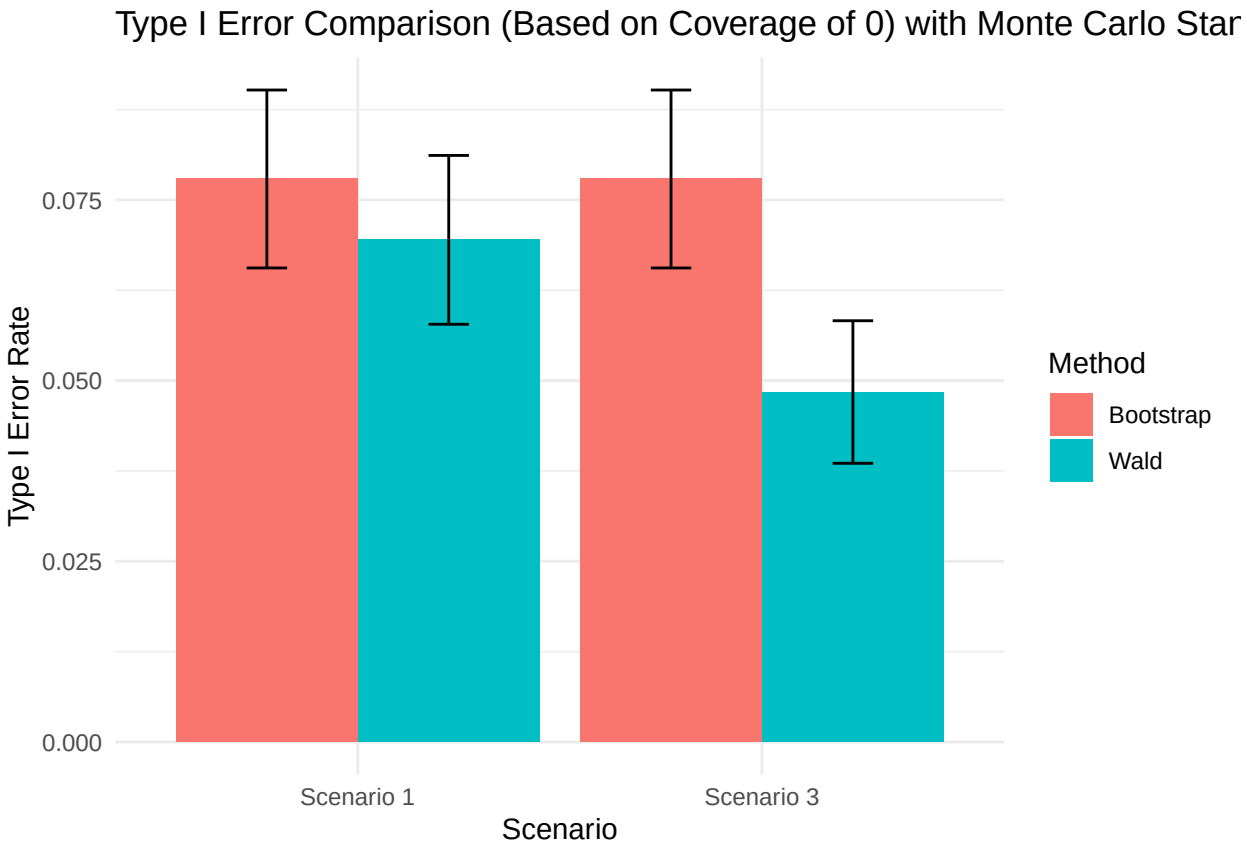
- Power

Under alternative assumption $\beta_{treatment} \neq 0$, simulation scenarios 2 and 4, we have:



- Type 1 error

Under null assumption $\beta_{treatment} = 0$, simulation scenarios 1 and 3, we have:



Write 1-2 paragraphs summarizing these results.

Based on the results above we have following results:

Estimates from normal error are almost non-bias, while we tend to overestimate under gamma error, because gamma random errors are always positive. Meanwhile, the standard error in normal is larger than gamma, which is because gamma with shape 1 rate 2 has a smaller variance than normal N(0,2). Wald tend to have a better coverage than Bootstrap under the 4 scenarios. In terms of hypothesis testing, the power of both methods under normal error have a much better power compared to gamma error scenario. Only under gamma error the wald estimate provide a type I error around 0.05, while all others are over 0.05, bootstrap is getting a larger type I error compared to wald uder both scenarios.