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.

Repo link: https://github.com/dliao1/bios731_hw3_liao

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
- Type 1 error

Write 1-2 paragraphs summarizing these results.

Table 2: Bias Summary Table

N	True Beta	Gamma	Normal
20	0.0	0.0118	-0.0122
20	0.5	0.0118	-0.0122

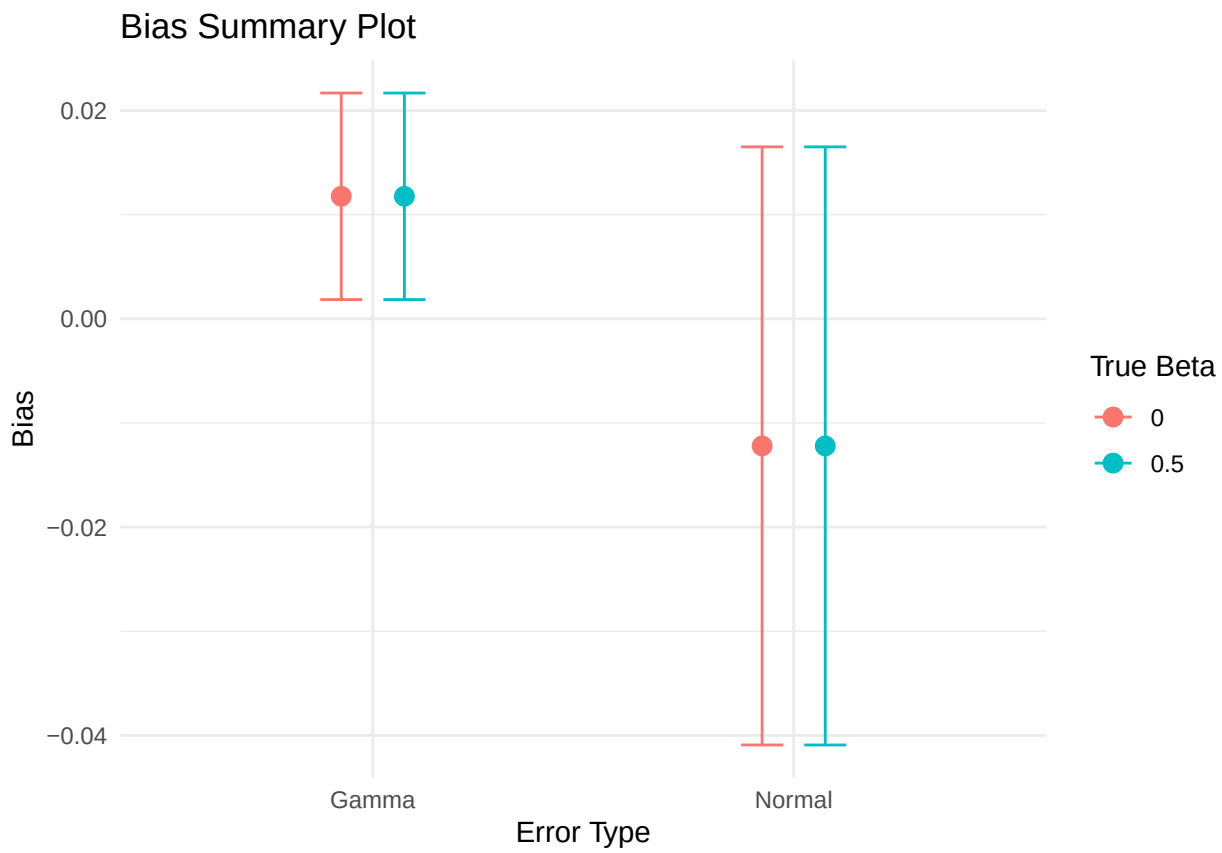
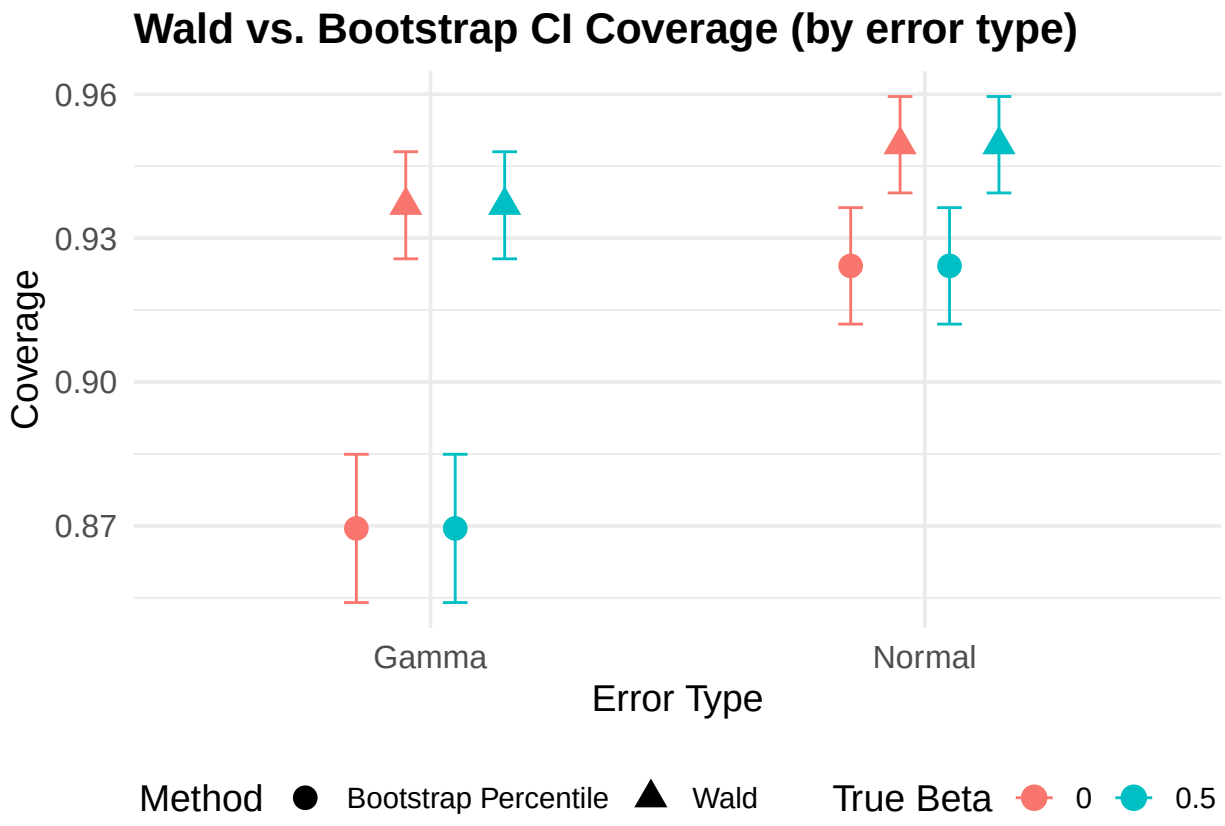
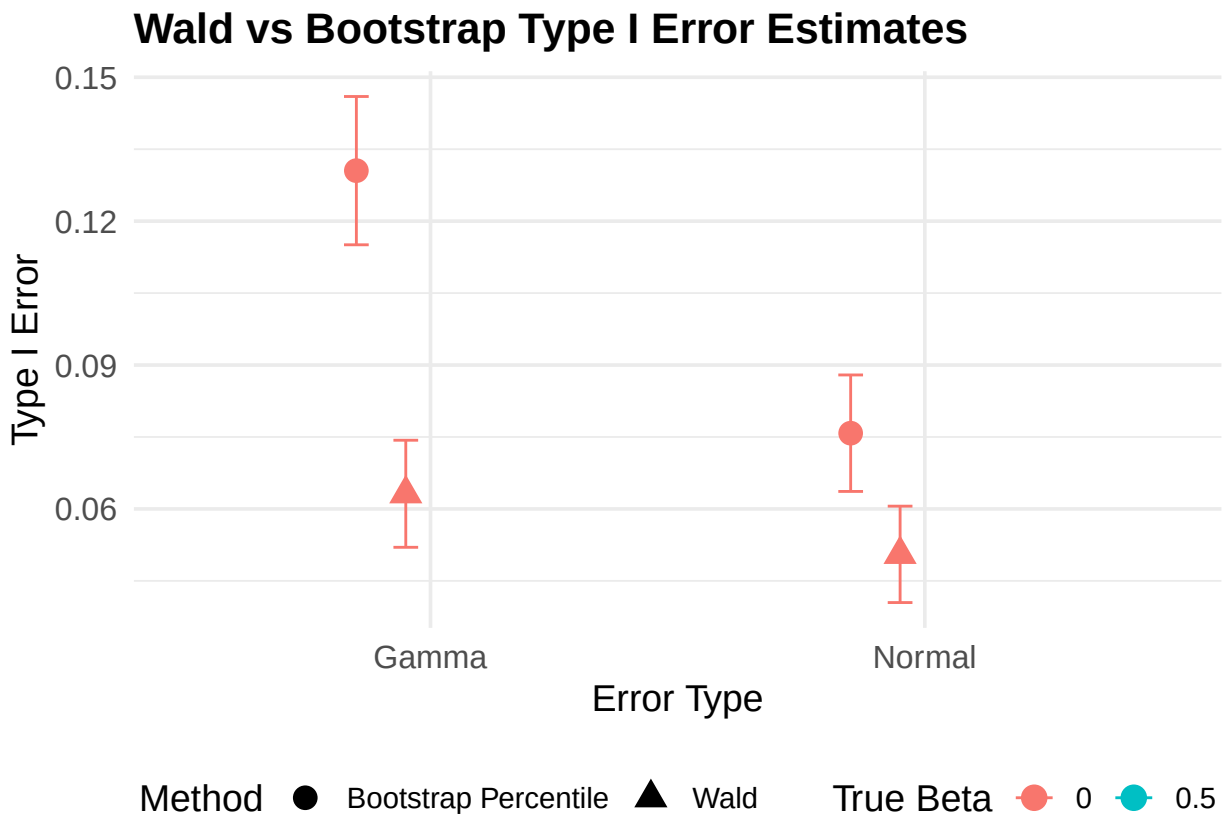
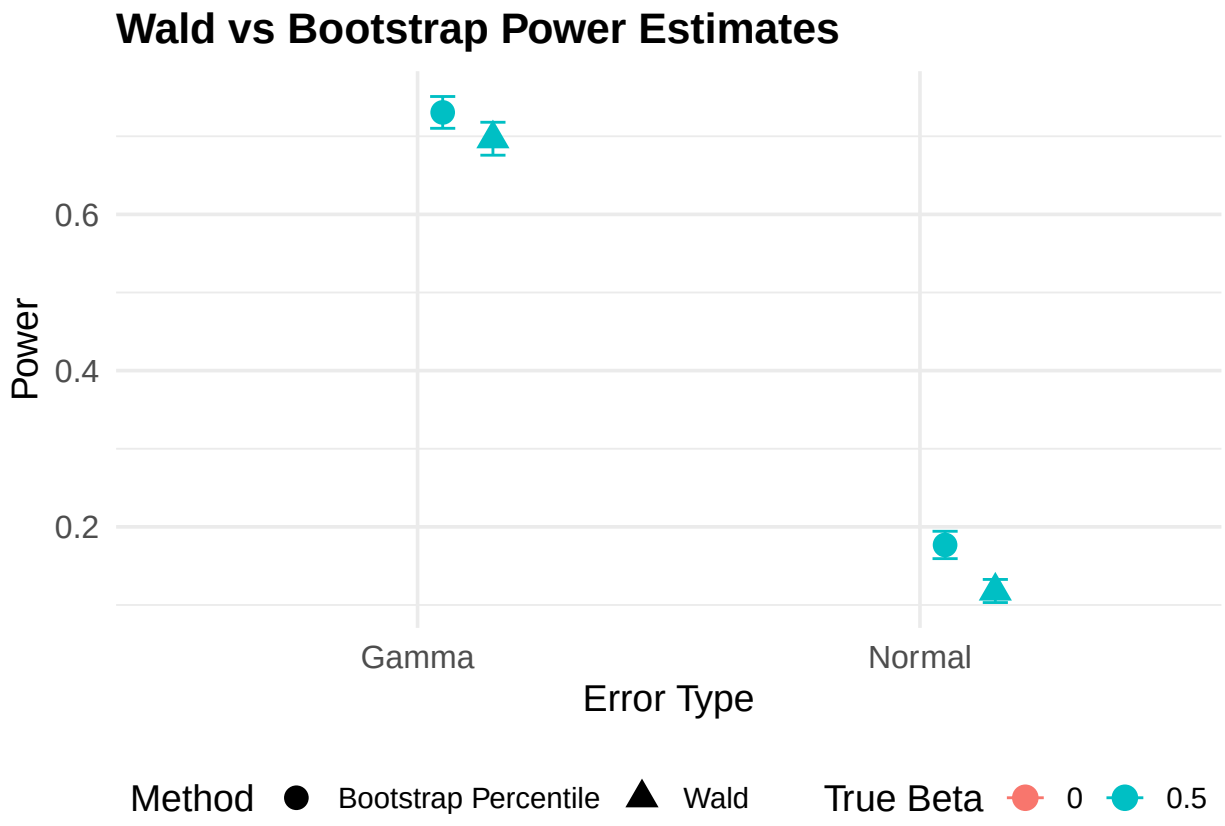


Table 3: Coverage Summary Table

		Gamma		Normal	
N	True Beta	Gamma Wald CI	Gamma Bootstrap Percentile CI	Normal Wald CI	Normal Bootstrap Percentile CI
20	0.0	0.937	0.869	0.949	0.924
20	0.5	0.937	0.869	0.949	0.924


Wald vs. Bootstrap CI Coverage (by error type)
Coverage
0.96
0.93
0.90
0.87
Gamma
Normal
Error Type
Method
Bootstrap Percentile
Wald
True Beta
0
0.5


Wald vs Bootstrap Type I Error Estimates
0.15
0.12
0.09
0.06
Type I Error
Gamma
Normal
Error Type
Method
Bootstrap Percentile
Wald
True Beta
0
0.5



Summary

To summarize, it looks like the bias for normal errors tended to be more negative, at around 0.012, while bias for gamma errors tended to be positive, at around -0.012. The Monte Carlo standard errors for normal errors looked to be larger than the MCSE for gamma errors, which makes sense, and these conclusions make sense when considering that the gamma distribution takes on only positive values, and that the variance of a gamma distribution with shape = 1 and rate = 2 is a lot smaller than the variance of 2 that we used for the normal distribution.

For coverages, Wald confidence intervals tended to have higher coverage for both true betas of 0 and 1 across both gamma and normal errors.

As for type I error and power, it looks like bootstrap percentile intervals had higher type I error rates than the wald intervals. For some reason, both bootstrap percentile intervals and wald intervals had really low power for both true betas of 0 and 0.5 for normal errors, but much higher power for gamma errors. My hypothesis was that Wald intervals would have higher power than the bootstrap intervals for the normal errors especially, but maybe due to the original sample size being n = 20 and the variance being higher than that of the gamma distribution, that was why the power was so low for the normal errors.