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.

Link to repo: https://github.com/dliao1/bios731_hw1_liao

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?

We will be running 3 * 3 * 2 = 18 simulation scenarios.

- What are the estimand(s)

The estimands are the average treatment effect $\beta_{treatment}$ and the standard error of treatment effect $se(\beta_{treatment})$.

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

3 methods are being compared - Wald confidence intervals, nonparametric bootstrap percentile intervals, and nonparametric boostrap t-intervals.

- What are the performance measure(s)?

The performance measures we are using 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.

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

We should perform 475 simulations for each simulation 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

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.

Bias

Table 2: Bias Summary Table

N	True Beta	Lognormal	Normal
10	0.0	-0.051	-0.093
10	0.5	-0.051	-0.093
10	2.0	-0.051	-0.093
50	0.0	-0.024	-0.006
50	0.5	-0.024	-0.006
50	2.0	-0.024	-0.006
500	0.0	-0.009	-0.009
500	0.5	-0.009	-0.009
500	2.0	-0.009	-0.009

This table shows the average bias for linear regression models fit at n = (10,50,500), true betas of (0, 0.5, 2.0), with errors either from the lognormal distribution (with mean 0 and variance log(2)) or the normal distribution (with mean 0 and variance 2).

Coverage

Table 3: Coverage Summary Table

		Lognormal			Normal		
N	True Beta	Lognormal Wald CI	Lognormal Bootstrap Percentile CI	Lognormal Bootstrap t CI	Normal Wald CI	Normal Bootstrap Percentile CI	Normal Bootstrap t CI
10	0.0	95.368	82.737	53.263	94.105	83.368	54.947
10	0.5	95.368	82.737	53.263	94.105	83.368	54.947
10	2.0	95.368	82.737	53.263	94.105	83.368	54.947
50	0.0	95.368	89.263	88.842	94.316	90.316	91.579
50	0.5	95.368	89.263	88.842	94.316	90.316	91.579
50	2.0	95.368	89.263	88.842	94.316	90.316	91.579
500	0.0	93.474	90.737	92.000	92.842	89.895	92.211
500	0.5	93.474	90.737	92.000	92.842	89.895	92.211
500	2.0	93.474	90.737	92.000	92.842	89.895	92.211

This table shows the coverage for confidence intervals calculated using 3 methods: Wald, Bootstrap Percentile, and Bootstrap t. Note that the number of bootstrap samples per simulation run was 50, with 10 nested bootstrap samples in the case of bootstrap t intervals.

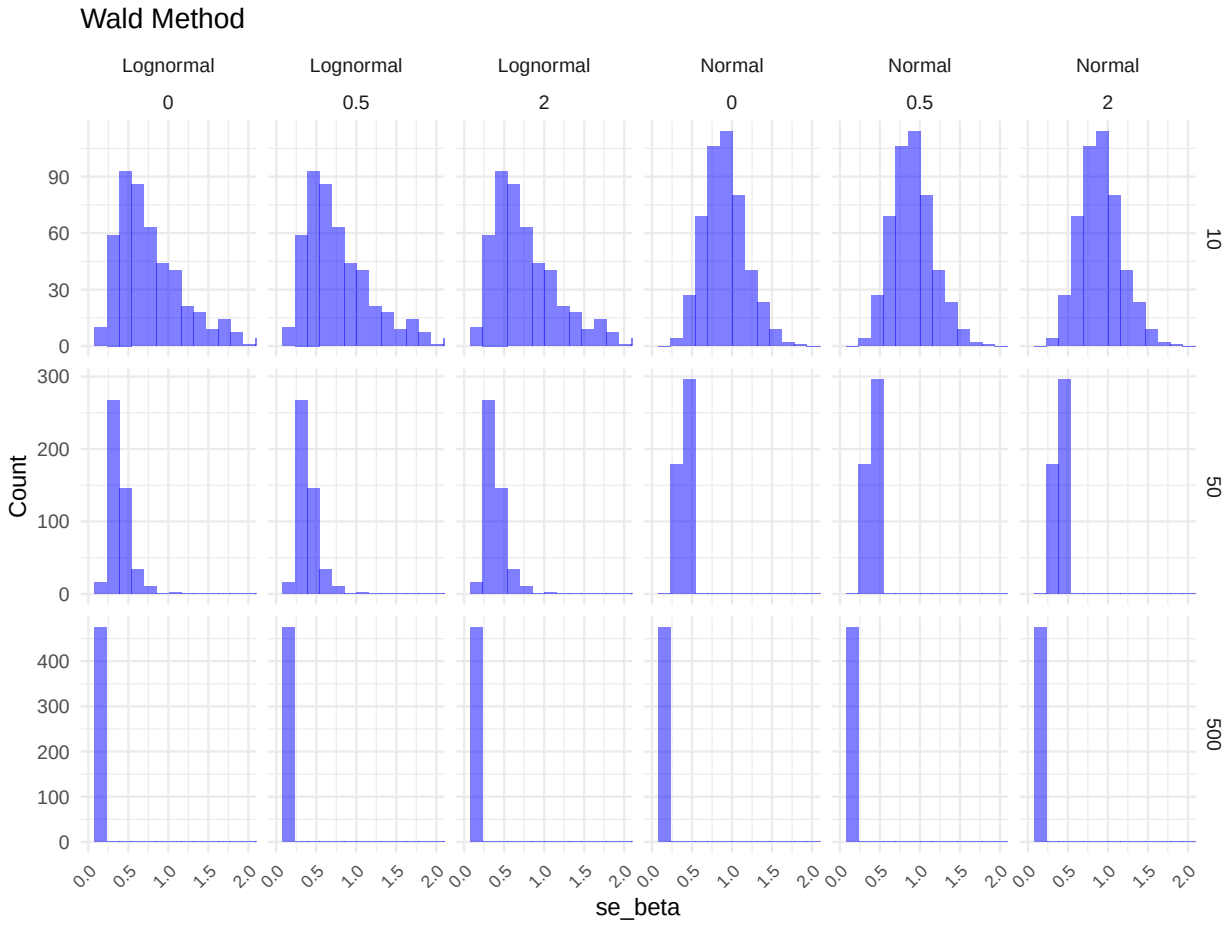
Computation Time

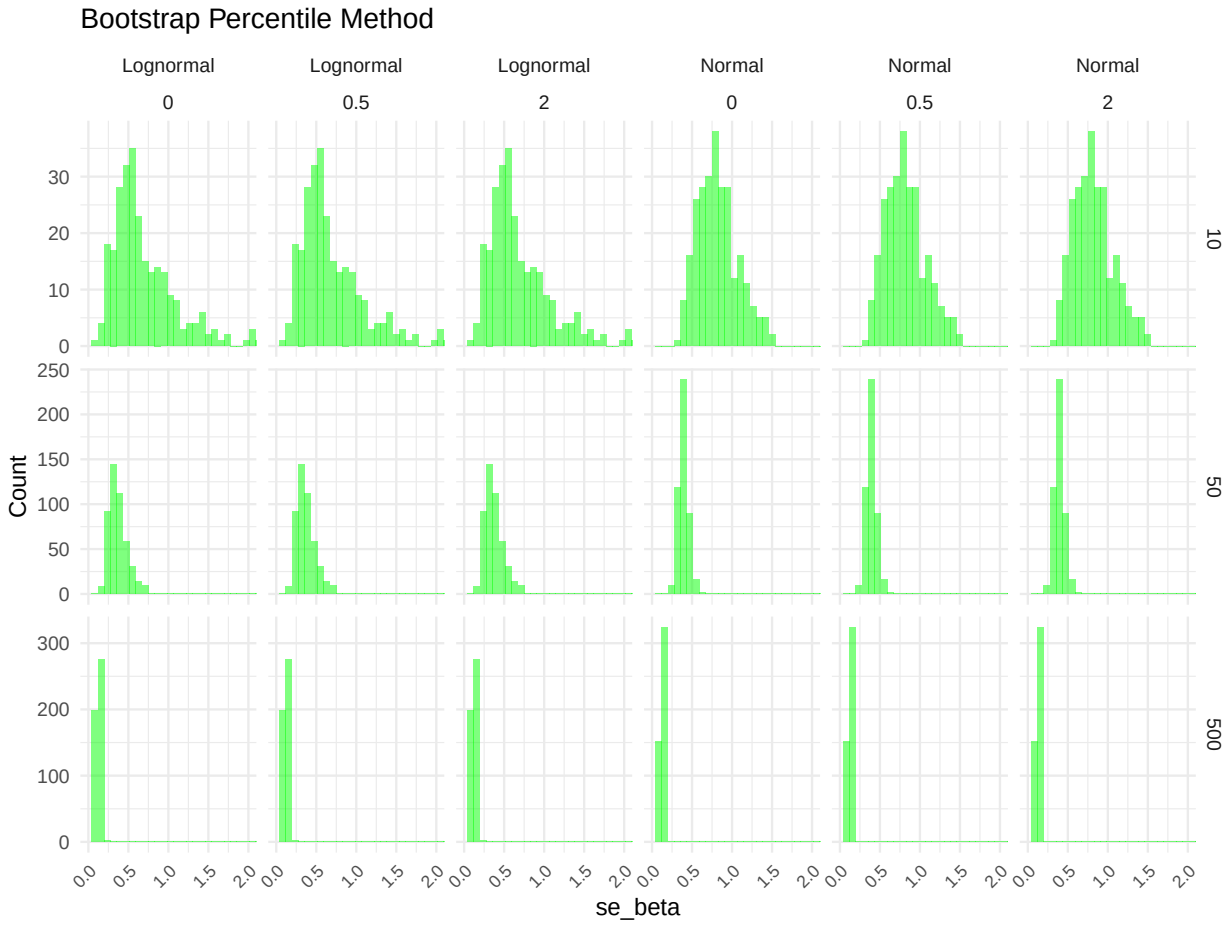
Table 4: Computation Time Summary Table

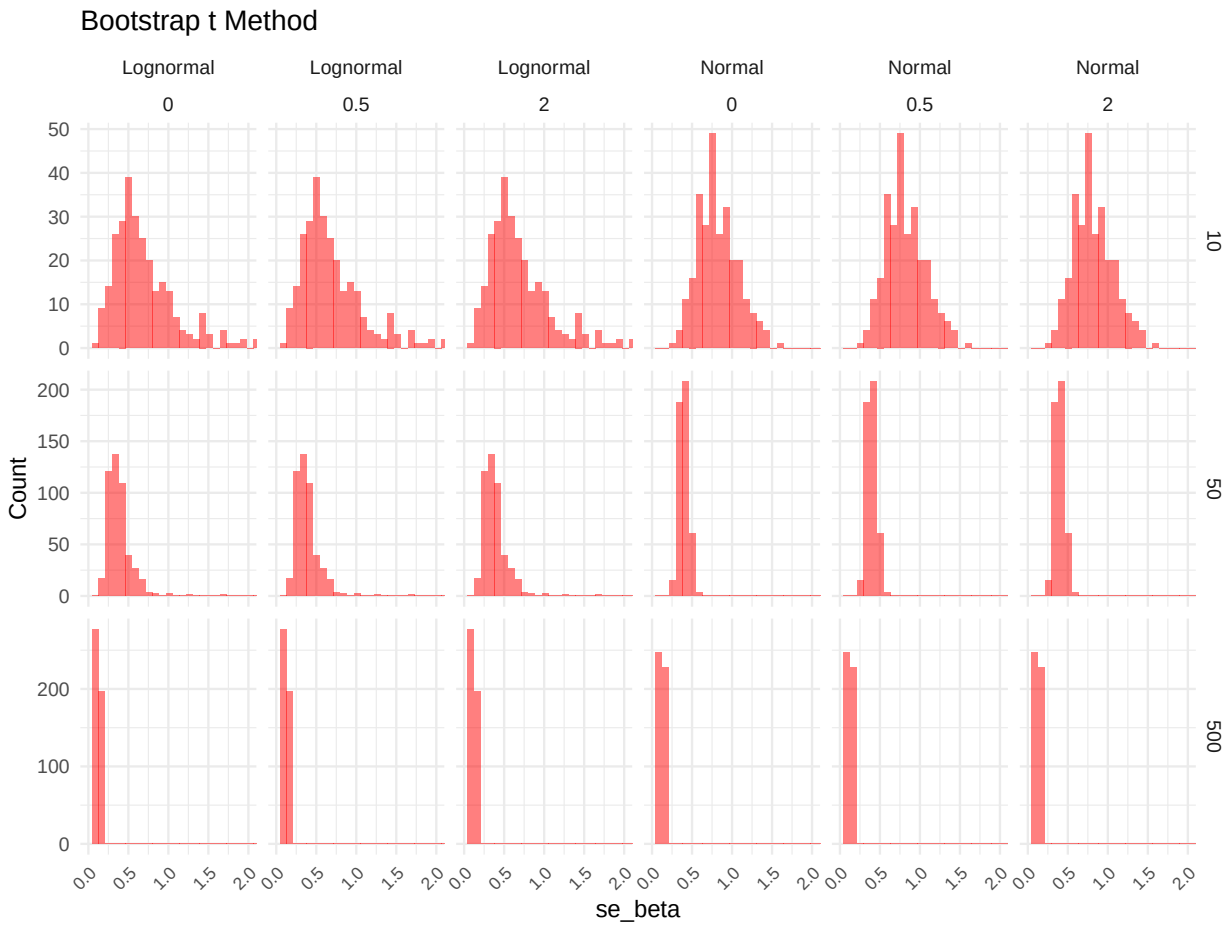
		Lognormal			Normal		
N	True Beta	Lognormal Wald Time	Lognormal Bootstrap Percentile Time	Lognormal Bootstrap t Time	Normal Wald Time	Normal Bootstrap Percentile Time	Normal Bootstrap t Time
10	0.0	0.030	1.669	18.328	0.025	1.440	15.947
10	0.5	0.024	1.419	15.662	0.025	1.444	15.941
10	2.0	0.025	1.453	16.050	0.025	1.437	15.828
50	0.0	0.024	1.415	15.462	0.025	1.437	15.723
50	0.5	0.024	1.420	15.550	0.025	1.432	15.639
50	2.0	0.025	1.441	15.814	0.025	1.436	15.709
500	0.0	0.024	1.423	15.645	0.025	1.458	15.944
500	0.5	0.025	1.446	15.876	0.026	1.447	15.886
500	2.0	0.025	1.467	16.014	0.024	1.460	16.005

This table shows the average computation time for confidence intervals calculated using 3 methods: Wald, Bootstrap Percentile, and Bootstrap t. Note that the number of bootstrap samples per simulation run was 50, with 10 nested bootstrap samples when calculating bootstrap t intervals.

Standard Error of Beta (Distribution)


Wald Method
Lognormal
Lognormal
Lognormal
Normal
Normal
Normal
0
0.5
2
0
0.5
2
10
50
500
Count
se_beta


Bootstrap Percentile Method
Lognormal
Lognormal
Lognormal
Normal
Normal
Normal
0
0.5
2
0
0.5
2
10
50
500
Count
se_beta



These three graphs show the distribution of standard errors for our estimated beta hat across n = (10, 50, 500) and true betas of (0, 0.5, 2), separated by distribution type (either lognormal or normal).

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.

The main findings of the simulation study regarding bias was that beta hats estimated from a distribution with right-skewed/lognormal errors tended to have a larger bias, but this bias seemed to decrease as sample size increased from 50 to 500, suggesting larger sample sizes could be used to "make up" for not having normally distributed errors, which is a key assumption when fitting linear regression models. Additionally, after computing coverages for the 3 methods used to generate confidence intervals (Wald, Bootstrap percentile, and Bootstrap t), it was found that Wald intervals tended to have the best overall coverage across all combinations of n, true betas, and errors generated from lognormal/normal distributions, at around ~93%. Comparing and contrasting the two bootstrap methods, the bootstrap t intervals tended to have better coverage than the bootstrap percentile intervals, especially when errors were non-normal, but also looked to have worse coverage at smaller sample sizes (when n = 10). However, since only 10 nested bootstrap samples were taken to calculate a t critical value, this poor coverage could also be in part due to this limitation.

- How do the different methods for constructing confidence intervals compare in terms of computation time?

In terms of computation time, Wald confidence intervals were the fastest to compute, taking on average only 0.03 seconds, while bootstrap percentile intervals took an average of 1.5 seconds. Bootstrap t intervals took the longest to calculate by far, averaging 16 seconds.

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

When errors were normally distributed, Wald intervals had the best coverage.

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

When errors were right skewed, Wald intervals still had the best coverage.