Homework 1

Sekha Daluwatumulle

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?

I would be running 18 different simulation scenarios.

- What are the estimand(s)

$\beta_{treatment}$, standard error of $\hat{\beta}_{treatment}$

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

Estimation using the standard regression calculations, also known as the wald method

Estimation using the bootstrap percentile interval method

Estimation using the bootstrap t interval method

Note: For all the bootstrap t calculations I used the residual method

- What are the performance measure(s)?

bias, coverage probabilities for the three methods, and also the computational times for the three methods

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.

number of simulations=475

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.

Table 2: A summary of the obtained results. Bias is calculated under the Wald's method. Coverage probabilities were calculated under the three methods given in problem 1.1. Time column indicates the elapsed time in seconds to conduct all the 475 simulations. Results are organised according to the different data generation procedures. Here we generated data from two distributions N(0,2) and lognormal (0,log(2)), with $\beta_{treatment}$={0,05, 2}, and n={10,50, 500}

			$Bias$	$Coverage$			$Time$		
				$Walds$	$Boot\,p$	$Boot\,t$	$Walds$	$Boot\,p$	$Boot\,t$
$N(0,2)$	$\beta = 0$	$n = 10$	-0.0276	0.9516	0.8589	0.8695	0.081	73.539	3.6768939×10^4
		$n = 50$	-0.0071	0.9516	0.9411	0.9432	0.11	86.615	4.3311226×10^4
		$n = 500$	-0.0055	0.9326	0.9326	0.9242	0.108	98.054	4.8971445×10^4
	$\beta = 0.5$	$n = 10$	-0.0276	0.9516	0.8589	0.8695	0.073	73.235	3.6817003×10^4
		$n = 50$	-0.0071	0.9516	0.9411	0.9432	0.102	86.957	4.3586414×10^4
		$n = 500$	-0.0055	0.9326	0.9326	0.9242	0.094	97.993	4.8960403×10^4
	$\beta = 2$	$n = 10$	-0.0276	0.9516	0.8589	0.8695	0.085	73.935	3.688619×10^4
		$n = 50$	-0.0071	0.9516	0.9411	0.9432	0.095	86.9	4.3480314×10^4
		$n = 500$	-0.0055	0.9326	0.9326	0.9242	0.107	95.966	4.8032057×10^4
$logN(0, log(2))$	$\beta = 0$	$n = 10$	-0.22	0.9642	0.9074	0.9116	0.072	74.376	3.7103137×10^4
		$n = 50$	0.0849	0.9642	0.9537	0.9347	0.102	85.807	4.2927402×10^4
		$n = 500$	0.013	0.9368	0.9411	0.9347	0.114	92.826	4.652906×10^4
	$\beta = 0.5$	$n = 10$	-0.22	0.9642	0.9074	0.9116	0.079	73.888	3.7035065×10^4
		$n = 50$	0.0849	0.9642	0.9537	0.9347	0.102	85.485	4.2913225×10^4
		$n = 500$	0.013	0.9368	0.9411	0.9347	0.113	89.472	4.4795673×10^4
	$\beta = 2$	$n = 10$	-0.22	0.9642	0.9074	0.9116	0.08	73.813	3.6920723×10^4
		$n = 50$	0.0849	0.9642	0.9537	0.9347	0.068	61.698	3.0925318×10^4
		$n = 500$	0.013	0.9368	0.9411	0.9347	0.081	65.76	3.2887456×10^4

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 bias using the standard regression approach (Walds) decrease as sample size increase. The bias values do not seem to change across $\beta_{treatment}$. Coverage probabilities for Walds method are all closer to 95%. However, it is slightly different (around 96%) under the lognormal distribution. Contrary to the belief that when n is coverage probabilities are high coverage, we can see that coverage for Wald's when n=500 is around 93%. The coverage of both the bootstrap methods is low when n is small (n=10). It is around 85% and 90% under the normal and lognormal assumptions respectively. As sample size increase, both the boostrap methods perform well, with coverage probabilities close to 95%. Under the skewed data and When n=500, bootstrap percentile interval's coverage probability is the closest to 95%. This indicates that bootstrap method works better when data are skewed (data generated from lognormal) The coverage probabilities don't seem to change across $\beta_{treatment}$ values either.

By comparing Figure 1 and 2, we can see that the distributions of standard deviation of $\hat{\beta}_{treatment}$ for the two methods (Wald and bootstrap) are quite similar with slight changes. Under the lognormal data assumption, the range of $se(\hat{\beta}_{treatment})$ is slightly larger in Wald's method compared to bootstrap methods. When the data are generated using a normal distribution, the distribution of $se(\hat{\beta}_{treatment})$s also look like a normal distribution visually, and when the data are generated using the lognormal distribution, the distribution of $se(\hat{\beta}_{treatment})$ seems to be skewed as well. When the sample size increase, the range of $se(\hat{\beta}_{treatment})$ decrease quite considerably, indicating that when n is high there is low variance in the estimates. The distribution of $se(\hat{\beta}_{treatment})$ do not seem to vary across different $\beta_{treatment}$

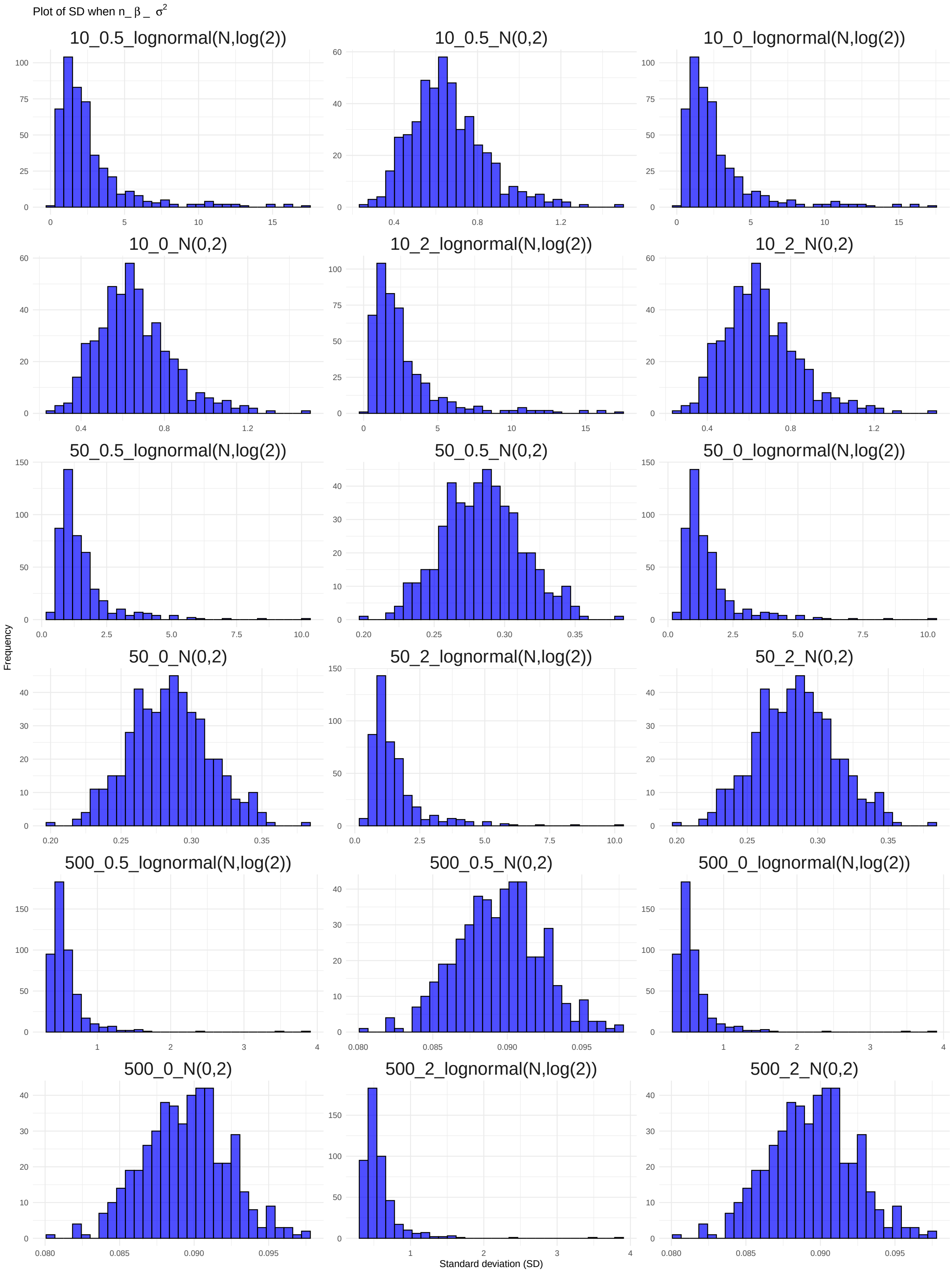

Figure 1: Distribution of standard deviation of estimated using walds method

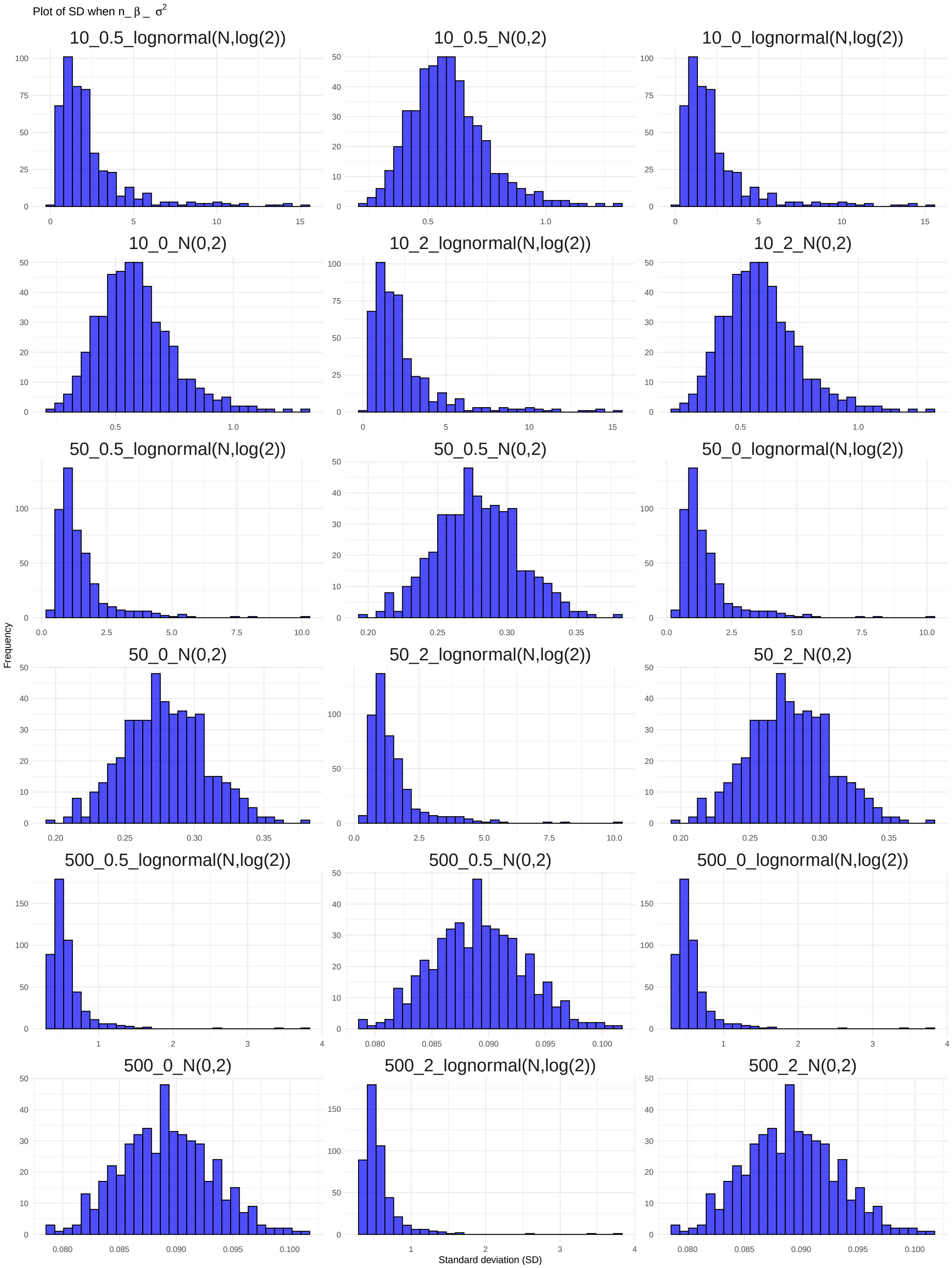

Figure 2: Distribution of standard deviation of estimated using bootstrap method

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

The computaional times calculated in Table 1 is the total elapsed time in seconds took to complete all the 475 simulations. We can see that Wald's method took the least number of seconds. Therefore, Wald's method is the fasted, next bootstrap percentile and the slowest is the boostrap t interval method.

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

Walds method

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

Bootstrap percentile when sample size is large enough, if the sample size is like 10 Wald would be better