Homework 3

GitHub Repo Link:

https://github.com/rw417/bios731_hw3_wan

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.

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:

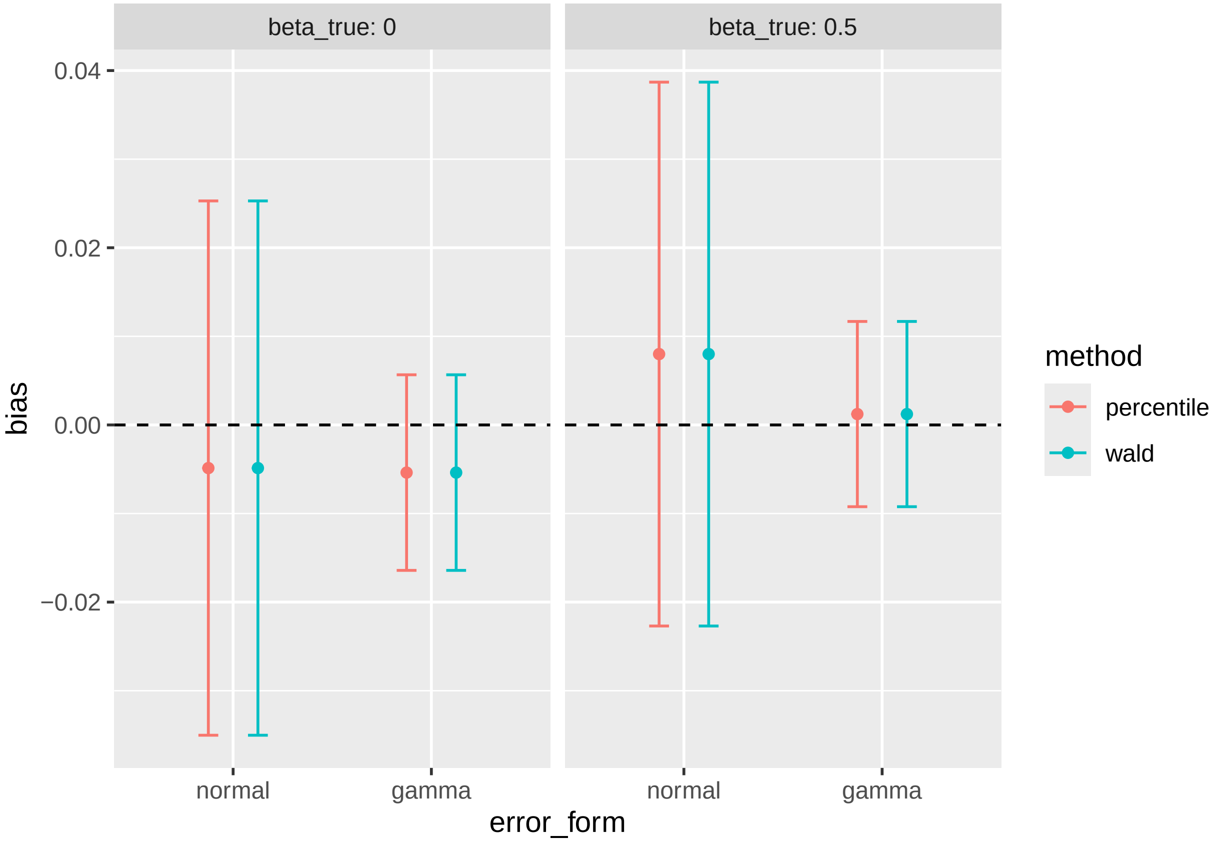
Workflow: * Use structured scripts and subfolders following guidance from the cluster computing project organization lecture * Instead of parallelizing your simulation scenarios (as in HW1), each simulation scenario should be assigned a different JOBID on the cluster.

Presenting results:

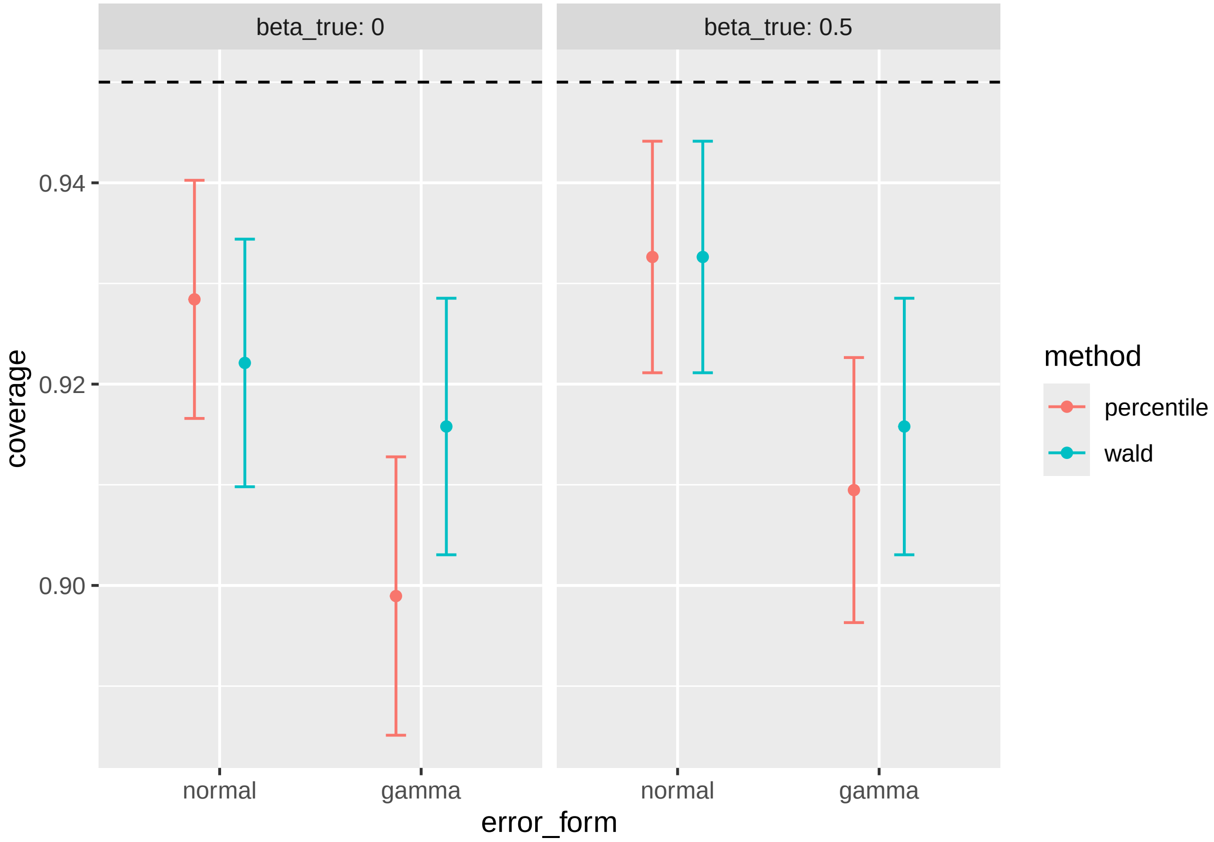
Create plots with *Monte Carlo standard error bars* to summarize the following:

- Bias of $\hat{\beta}$
- Coverage of $\hat{\beta}$
- Power
- Type 1 error

Write 1-2 paragraphs summarizing these results.


beta_true: 0
beta_true: 0.5
0.04
0.02
0.00
−0.02
bias
normal
gamma
normal
gamma
error_form
method
percentile
wald

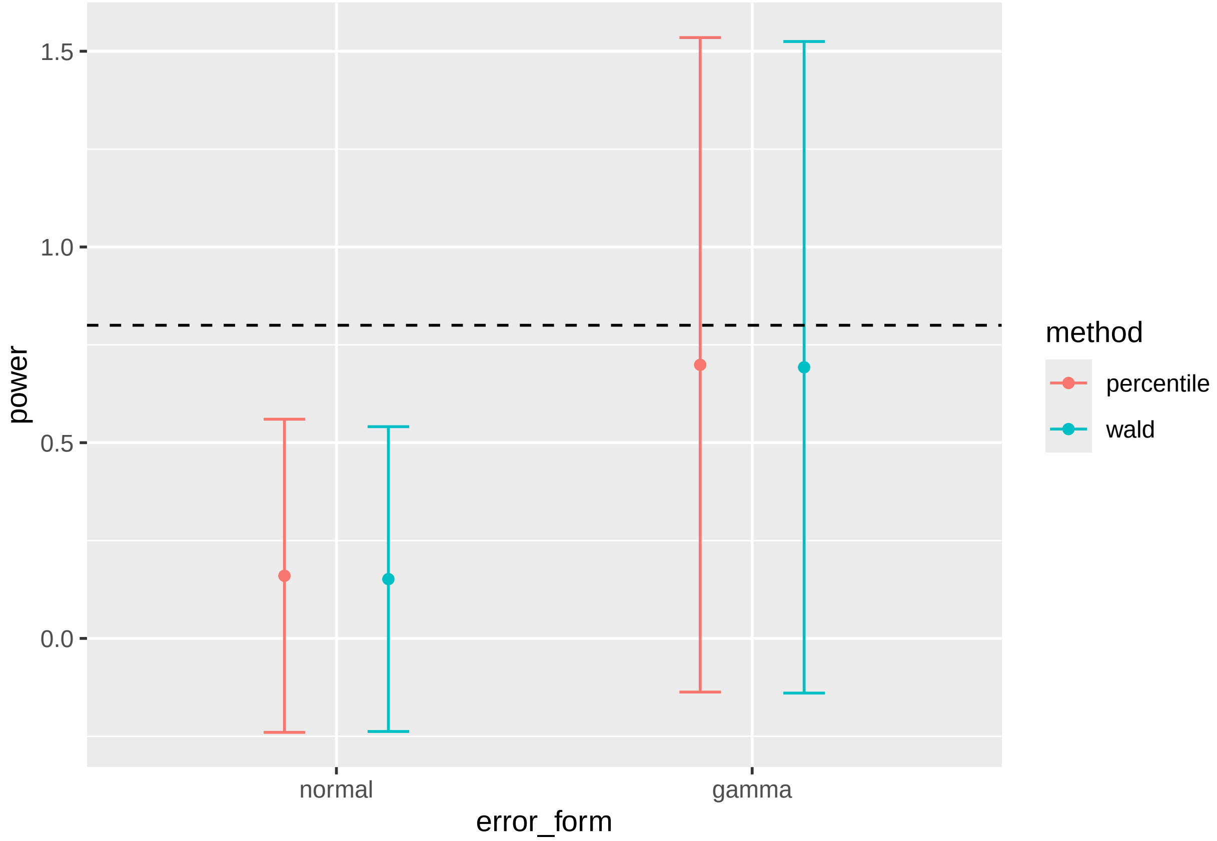
Bias


beta_true: 0
beta_true: 0.5
coverage
0.94
0.92
0.90
normal
gamma
normal
gamma
error_form
method
percentile
wald

Coverage

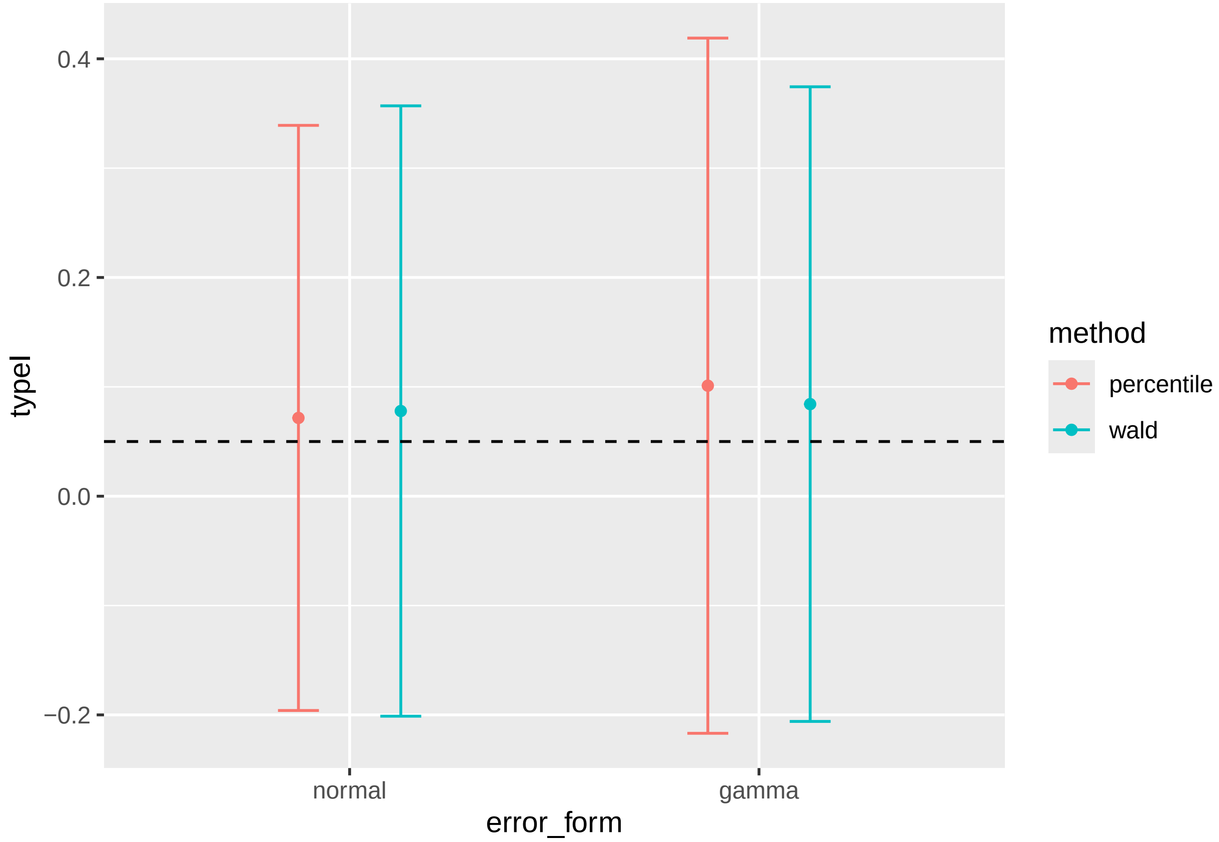
Power

We calculated the power for correctly rejecting $H_0 : \beta_{treatment} = 0$ when the truth is $\beta_{treatment} = 0.5$.

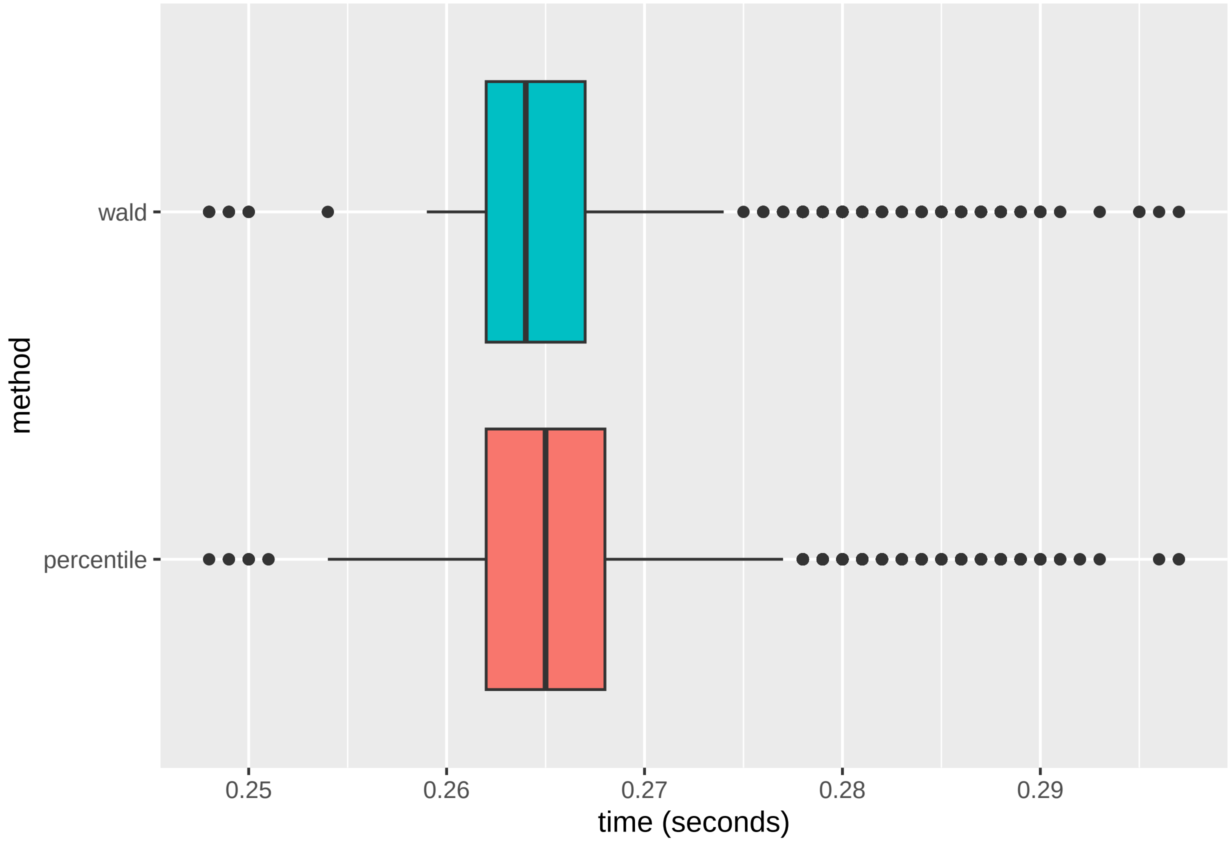


Type I error

We calculated the power for incorrectly rejecting $H_0 : \beta_{treatment} = 0$ when the truth is $\beta_{treatment} = 0$.



Computation Time



Summary

Both the Wald and the percentile methods produced the exact same bias estimates for $\hat{\beta}$ which is expected because we used the same bootstrap datasets for both methods. Both methods had higher coverage when we correctly specified the error distribution for the regression. When the error distribution was misspecified, the percentile method had higher coverage than the Wald method, suggesting that it is more robust. This advantage is more noticeable when the true β is 0. Unsurprisingly, both methods produced very low power estimates when we correctly specify the error distribution because our sample size is very small at $n = 20$ and the true effect size is not huge at $beta_{treatment} = 0.5$. However, power estimates were quite high when we misspecify the error distribution. This was probably because we tended to over estimate the effect size when the error has a gamma distribution (biases were positive when errors were gamma distributed). Type I error rates were well controlled for both methods under both error distributions. The standard error estimates for statistical power and Type I error were quite large. We would need to run more simulations to get more precise estimates. The percentile method took slightly longer to run than the Wald method, but the difference was not huge.