STAT 600 Statistical Computing

HW 1: Packages, Markdown, Parallelization

Spring 2024, **Due Feb. 1st**

Homework format: Homework should be submitted as a pdf generated by LaTeX or Rmarkdown. All functions should be coded in `Rcpp/RcppArmadillo`. Please provide explanations of your solutions and appropriate graphics (labeled well).

1. Make an `R` package called `SimpLin` for running a simple linear regression model that:
   - takes in numeric vectors $\boldsymbol{x}$ and $\boldsymbol{y}$
   - outputs estimated regression coefficients, $\hat{\beta}_0$ and $\hat{\beta}_1$, their corresponding standard errors and 95% confidence intervals, residuals, and predicted values as a list.
   - wraps the `cpp` function (`SimpLinCpp`) in an `R` function (`SimpLinR`) that throws errors if $\boldsymbol{x}$ and $\boldsymbol{y}$ are not numeric vectors of the same length
   - provides a description and brief vignette (using .Rmd file) demonstrating how to use the package
2. Connect and manage the development of your `R` package with your GitHub account (make one if you do not have one). Submit link in HW. I should be able to download and install your `R` package locally.
3. Simulate 100 data sets with $n = 100$ observations each, where $x \sim N(0, 1)$ and error terms $\epsilon \sim N(0, 1)$ with true regression coefficients $\beta_0 = 1$ and $\beta_1 = -1$. Fit a linear regression model to each of the data sets using your package in (1) and the `lm()` function in `R` **in parallel**. Calculate the runtime for each of the data sets using both models. (Note that you do not have to run in parallel using Rcpp.
4. Provide a table of summary statistics for the simulations including average runtime, bias, coverage probability (proportion of 95% CIs that include the true regression coefficients), mean squared error for regression coefficients, and predictive mean squared error for $\hat{y}$ across all simulations for your model and `lm()`. Plot a histogram of the estimated regression coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ across all simulations. Comment on the performance of the methods.