Comparison of Serial and Parallel Execution of Bootstrapped GLMs

Introduction

This report compares the results of serial and parallel execution of Generalized Linear Model (GLM) fitting on bootstrapped samples from the Boston dataset. The comparison is based on two key aspects:

1. The similarity or difference in the distributions of model fit statistics (AIC values).
2. The execution times for both approaches.

Note: The cores on my computer are 4, so the execution time may improve for computers having more cores.

Methodology

- **Data**: The Boston dataset from the `MASS` package in R.
- **Bootstrapping**: 100 bootstrapped samples were generated from the dataset.
- **GLM Fitting**: Each sample was used to fit a regression GLM with `medv` as the outcome variable.
- **Execution Approaches**:
  - **Serial Execution**: Models were fit sequentially in a for-loop.
  - **Parallel Execution**: Models were fit in parallel using the `foreach` and `doParallel` packages.
- **Comparison Metrics**:
  - Model fit statistics (AIC values)
  - Execution time (measured using `microbenchmark()`)

Comparison of Model Fit Statistics

The Akaike Information Criterion (AIC) was extracted from each model to assess fit.

Below are the Plots Created for the Serial Execution and Parallel Execution-

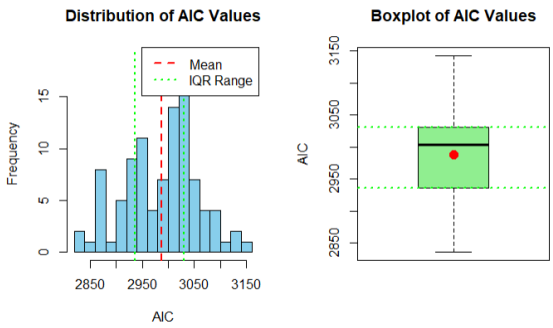

Fig 1. Serial Execution

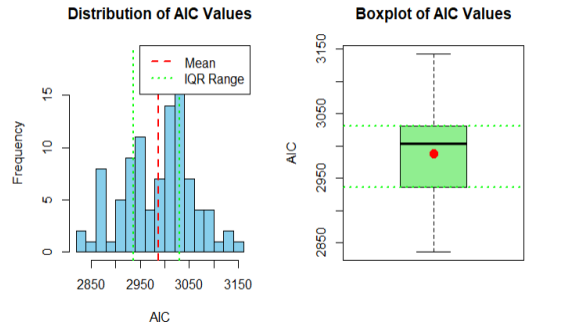

Fig 2. Parallel Execution

From the plot, we can see that they are identical.

Statistical Summary

Execution Method	Mean AIC	Inter-Quartile Range (IQR)
Serial	2987.725	94.02024
Parallel	2987.725	94.02024

Findings:

- The mean and IQR values for AIC in serial and parallel execution are nearly identical, indicating that parallel execution did not affect model estimation.
- The histograms of AIC values for both executions show similar shapes and spread, reinforcing this similarity.

Comparison of Execution Times

Execution time for both approaches was measured in seconds.

Execution Method	Time (milliseconds)
Serial	662.5326
Parallel	1144.2562

Findings:

- Surprisingly, parallel execution took longer than serial execution.
- The overhead of setting up and managing parallel processes likely outweighed the benefits of parallel computation.
- Parallelization is more effective for larger datasets or more computationally intensive models.

Conclusion

- **Model Fit Consistency**: The distributions of AIC values for serial and parallel executions are nearly identical, confirming that parallelization does not introduce significant differences in model estimation.
- **Execution Time Differences**: Parallel execution was slower due to overhead costs, indicating that for small tasks, serial execution remains more efficient.
- **Recommendation**: Parallel execution should be reserved for large-scale modeling tasks where the computational burden justifies the parallelization overhead.