**Project #1: Parallel Bootstrapping**

Bootstrapping is a method of statistical inference dependent on creating numerous random samples using replacement. An appropriate calculation is made on each sample, then all calculations are aggregated, and an inference is made. Whereas this procedure is very popular due to its accuracy and lack of assumptions, one major concern is its large computation time.

The goal of this project was to construct a confidence interval for each coefficient in a full fitted model using bootstrapping, all within 10 minutes. The dataset consisted of 11 independent variables for 100,000 observations which would take around 4 hours if ran sequentially (~0.15 sec/model). To expedite this to within 10 minutes, a procedure like Map/Reduce was used. The bootstrapping task was split into 25 jobs, equating to 4,000 resampled regressions per job. Each job was then submitted using slurm scripts to execute the R code on Seawulf. Once completed, 25 .csv files were outputted in the home directory, each consisting of 4,000 estimates for each coefficient. All these estimates were aggregated to then be reduced to each coefficient’s 2.5 and 97.5% quantiles.

DATA

Job 25

(4,000x)

Sample data

Fit model

(4000x)

Job 2

Job 1

(4,000x)

Sample data

Fit model

(4000x)

(4,000x)

Sample data

Fit model

(4000x)

Reduce

Combine all coef. estimations

Calculate 2.5 + 97.5% quantile per coef.

Figure 1: Project outline

Regression theory states that, under Gauss-Markov Assumptions, the coefficients have an asymptotically sampling distribution with support of the central limit theorem. Furthermore, without the assumption of the normality of errors, under a large sample size, a linear regression’s coefficients will follow asymptotic normality. This can be represented mathematically as:

With this assumed probability distribution various inferences can easily be calculated, including confidence intervals.

|  |  |  |
| --- | --- | --- |
| **Variables** | **95% Bootstrap CI** | **95% Asymptotic CI** |
| Intercept | (2.942, 3.066) | (2.942, 3.066) |
|  | (1.129, 1.252) | (1.128, 1.252) |
|  | (0.093, 0.113) | (0.093, 0.113) |
|  | (-0.024, 0.100) | (-0.025, 0.099) |
|  | (-0.020, 0.104) | (-0.020, 0.104) |
|  | (0.558, 0.682) | (0.558, 0.682) |
|  | (0.356, 0.479) | (0.356, 0.480) |
|  | (-0.024, 0.100) | (-0.023, 0.100) |
|  | (0.978, 1.022) | (0.980, 1.023) |
|  | (-0.039, 0.085) | (-0.038, 0.086) |
|  | (-0.092, 0.032) | (-0.092, 0.032) |

Table 1: Bootstrap and asymptotic confidence intervals

In *Table 1,* we see that the 95% confidence intervals for the Bootstrap and Asymptotic calculations are extremely close. However, there are several cases in which the margin of error for the bootstrap interval is slightly wider than that of the asymptotic interval. With this in mind, the interval from asymptotic calculation would be preferred for inference due to its preciseness.

For this project it was seen that, although the intervals were almost identical, the asymptotic interval had the slight advantage for this data set. However, this is not uniform for all cases. When using a smaller dataset, the assumption of a sufficiently large size may not be adequate for coefficients to be asymptotically normal. In addition, another advantage to bootstrapping is that it does not require the Gauss-Markov assumptions to be valid. But when dealing with a large sample we may prefer to use an asymptotic distribution due it having more precision and significantly shorter computation time.