AMS 598 Project 1 Report

Jiecheng Song

September 2021

1 Introduction

In project1_data.csv, there is one response variable y and ten explanatory variables (X1-10). Run a linear regression (y $\sim$ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10) and use the bootstrap method (B=100,000 times) to construct confidence intervals of the coefficients.

Note that it roughly takes 0.15 second to run one linear regression with the given data, so the whole bootstrap procedure takes 15,000 second (4 hours) if it is run sequentially.

Use the concept of Map/Reduce to speed up the computation and implement it with SeaWulf. The goal is to complete the bootstrap procedure within 10 minutes.

Compare your results with asymptotic confidence intervals derived from regression theory and write a report about your analyses. Please submit both report and scripts to Blackboard.

2 Methods

To finish bootstrap in 10 mins, we should separate the main job in 25 jobs and 4000 iterations in each job, which will use about 10 mins for each node.

There are three steps in this project: 1.file generate 2.write coefficients 3.analysis

2.1 File generate

Use one R file and slurm file to generate 25 (file_generate.R, file_generate.slurm) to create 25 separate R files and slurm files (use write function) to write coefficients into 25 different files.

2.2 Write coefficients

Use sample function to do bootstrap re-sampling, repeat 4000 iterations in each file, submit 25 files to computation nodes and write coefficients of regression into .csv files. (coef{i}.csv)

2.3 Analysis

Compare the empirical quantiles generated by bootstrap vs the theoretical confidence interval calculated by confint() function and calculate by regression formula. (the confidence interval by confint() function should be the same with calculated by formula).

3 Results and Analysis

The coefficients with original data set is:

coefficients	Estimation	Standard error	coefficients	Estimation	Standard error
Intercept	3.0041	0.0316	X6	0.4181	0.0315
X1	1.1903	0.0315	X7	0.0386	0.0316
X2	0.1029	0.0050	X8	1.0408	0.0315
X3	0.0372	0.0316	X9	0.0236	0.0316
X4	0.0422	0.0315	X10	-0.0302	0.0316
X5	0.6199	0.0317			

95% Confidence interval comparison:

Theoretical	2.5%	97.5%	Bootstrap	2.5%	diff	97.5%	diff
Intercept	2.9423	3.0660	Intercept	2.9423	0.0000	3.0659	-0.0001
X1	1.1285	1.2521	X1	1.1293	0.0008	1.2516	-0.0005
X2	0.0931	0.1127	X2	0.0931	0.0000	0.1127	0.0000
X3	-0.0247	0.0991	X3	-0.0241	0.0006	0.0989	-0.0002
X4	-0.0196	0.1040	X4	-0.0198	-0.0002	0.1041	0.0001
X5	0.5578	0.6820	X5	0.5577	-0.0001	0.6818	-0.0002
X6	0.3563	0.4799	X6	0.3568	0.0004	0.4792	-0.0008
X7	-0.0234	0.1005	X7	-0.0236	-0.0002	0.1007	0.0002
X8	0.9790	1.1026	X8	0.9788	-0.0002	1.1021	-0.0006
X9	-0.0383	0.0856	X9	-0.0387	-0.0003	0.0855	-0.0001
X10	-0.0921	0.0317	X10	-0.0920	0.0001	0.0318	0.0001

90% Confidence interval comparison:

Theoretical	5%	95%	Bootstrap	5%	diff	95%	diff
Intercept	2.9522	3.0561	Intercept	2.9524	0.0001	3.0559	-0.0002
X1	1.1384	1.2422	X1	1.1389	0.0005	1.2418	-0.0004
X2	0.0947	0.1111	X2	0.0947	0.0000	0.1112	0.0000
X3	-0.0147	0.0891	X3	-0.0144	0.0003	0.0889	-0.0003
X4	-0.0097	0.0941	X4	-0.0100	-0.0004	0.0942	0.0001
X5	0.5678	0.6721	X5	0.5678	0.0000	0.6719	-0.0002
X6	0.3663	0.4700	X6	0.3667	0.0005	0.4695	-0.0005
X7	-0.0134	0.0905	X7	-0.0135	-0.0001	0.0908	0.0002
X8	0.9889	1.0927	X8	0.9886	-0.0004	1.0924	-0.0003
X9	-0.0284	0.0756	X9	-0.0287	-0.0003	0.0758	0.0001
X10	-0.0821	0.0217	X10	-0.0818	0.0003	0.0219	0.0001

According Central Limit Theorem, with sufficiently large sample size, we could consider the result of bootstrap following a normal distribution. Then the qth sample quantile following a normal distribution with mean is the theoretical qth quantile x_q and varaince is $q(1-q)/(nf_x(x_q)^2)$. [1]

So the variance for 95% CI bound should be $0.025*0.975/(100000*(f_{normal}(Z_{0.025})^2) \approx 7.1359e-05$ and standard deviation about 0.0084 and for 90% CI bound should be $0.05*0.95/(100000*(f_{normal}(Z_{0.05})^2) \approx 4.4656e-05$ and standard deviation 0.0066. From above table, the difference between bootstrap value and theoretical value are all less than 0.01, which is less than one standard deviation, which proves our bootstrap simulation succeeded.

[1] get from: https://stats.stackexchange.com/questions/14877/central-limit-theorem-for-sample-quantiles (Statistics and Data Analysis for Financial Engineering, Ruppert, David)