ordinary-least-squares application

Caleb Yenusah

June 2022

1 Introduction

The *ordinary-least-squares* application calculates the ordinary least squares(OLS) estimates using the the following relation:

$$\beta = (X^T X)^{-1} X^T y \tag{1}$$

2 Application design

The design process of the application is as follows:

- Method 1: computes the OLS estimates directly by using the Normal Equation $\beta = (X^T X)^{-1} X^T y$. This method is referred to as the NE method hereafter.
- Method 2: computes the OLS estimates by first rearranging the Normal Equation as $(X^T X)\beta = X^T y$ and then solve the systems of linear equations $Ax = b$ using the LU factorization method. This method is referred to as the AxbLU method hereafter.
- The application uses CBLAS and LAPACKE libraries for the linear algebra operations. This enables easy linkage of the application with more optimized and parallel versions of these libraries such as Intel Math Kernel Library.
- The application uses the CMake build system.

3 Dependencies

The application uses CBLAS and LAPACKE linear algebra libraries. The libraries can be installed using the following command:

```
$ sudo apt install libopenblas-dev
$ sudo apt-get install liblapacke-dev
```

4 Basic build and usage

After cloning the project use the following commands to build it.

```
$ cd ordinary-least-squares
$ mkdir build
$ cmake -S . -B build
$ cmake --build build
```

This builds the executable estOLS in /build/ordinary-least-squares.
The program accepts command line arguments which can be probed using the following command:

```
$ estOLS --help
```

The above command shows the following output:

```
Arguments are:
  long      short  arg        description
--Xmat        -x    1    X matrix csv file (full path)
--yVec        -y    1    y Vector csv file (full path)
--XmatSkip    -s    1    number of header lines to skip in Xmat file (default 0)
--yVecSkip    -k    1    number of header lines to skip in yVec file (default 0)
--method      -m    1    select which method to use 0, 1 (default 1, see below for more info)
--writeFile   -w    0    output result to file (default writes result to screen)
--benchmark   -b    1    run benchmark for Xmat dimensions in file
--help        -h    0    print this message
More info:    method 0: compute OLS estimates using direct evaluation of the normal equation
  method 1: compute OLS estimates using using LU factorization to solve Ax=b
  Note: options with arg=1 require values.
```

If X matrix and y vector are stored in a csv file named XMAT and yVEC with no header lines, the program can be run with the following minimal commands:

```
$ estOLS -x $path_to_file/XMAT.csv
  -y $path_to_file/yVEC.csv
```

Example of XMAT.csv and yVEC.csv are included in the test folder. To run the program with the files, you would need to specify the number of header lines in each file using the -s and -k options for the XMAT.csv and yVEC.csv files, respectively. To specify that the program writes the result to an output file (OLSest.csv) use the -w flag. Below is the full command:

```
$ estOLS -x $path_to_file/XMAT.csv
  -y $path_to_file/yVEC.csv -s 1 -k 1 -w
```

5 Benchmark

The two methods for calculating the OLS estimates were benchmarked for X matrix $(n \times m)$ dimension with the n dimension set to 40000 and the m dimension set to [500,1000,2000,4000,10000]. The benchmark was ran three times and the results were averaged and plotted in the figure below. From the Figure 1, it can be observed that the AxLU method outperforms the NE method. For an X matrix dimension of 40000 × 10000, the AxLU method is 3.4 times faster than the NE method. The difference in performance can be attributed to the reduced number of operations in the AxLU method compared to the NE method. The benchmark was performed on a laptop powered by Intel core i5 8th Gen processor.

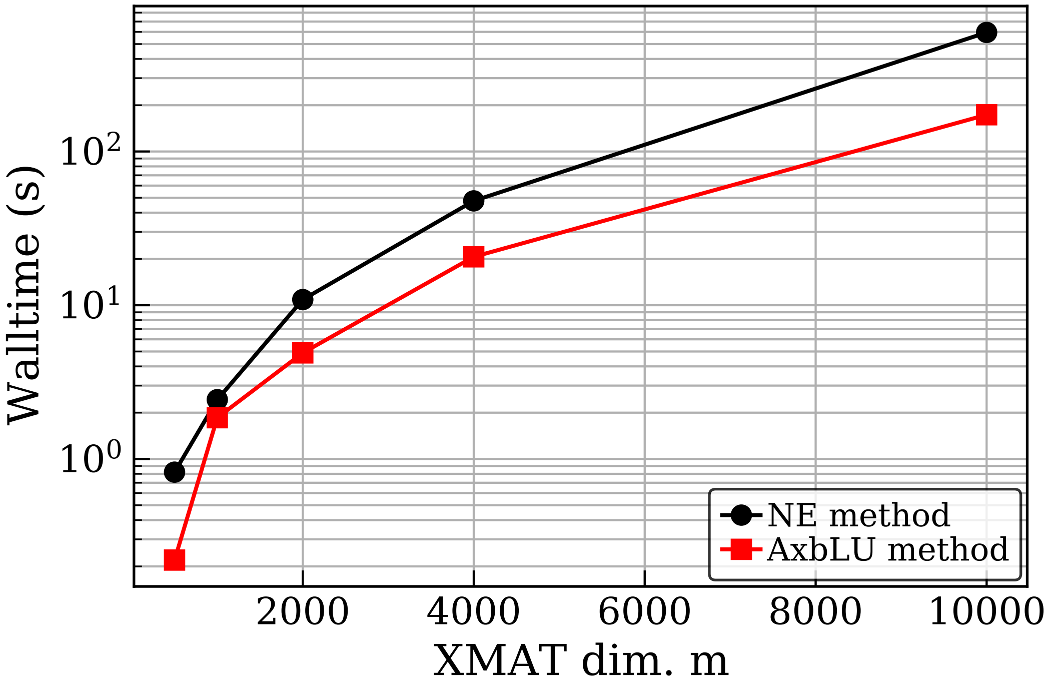

Figure 1: Benchmark results for computing the OLS estimates using the NE and AxbLU methods.

The benchmark option for the program accepts a csv file that lists all the n and m dimensions of the X matrix that should be benchmarked. An example of the benchmark file is shown in benchmark/benchmark.csv. The file can be modified to run different X matrix dimensions. The command to run the benchmark study is shown below:

```
$ estOLS -b $path_to_file/benchmark.csv
```

6 Future work

Future work would look at implementing an iterative method such as the conjugate gradient method to solve the linear systems of equations ($Ax = b$) to calculate the OLS estimates.