Lecture 2: Advanced R

Zhentao Shi

Feb 18, 2017

Advanced R

Introduction

In this lecture, we will talk about efficient computation in R.

- R is a vector-oriented language. In most case, vectorization speeds up computation.
- After optimizing a piece of code, we turn to more CPUs for parallel execution to save time.
- Clusters are accessed remotely. Communicating with a remote cluster is different from operating a local PC.

Vectorization

Despite mathematical equivalence, different ways of calculation can perform distinctively in terms of computational speed.

Does computational speed matter? For a job that takes less than a minutes, the time difference is not a big deal. For modern economic structural estimation problems commonly seen in industrial organization, a single estimation can consume up a week. Code optimization is essential there.

Other computational intensive procedures include bootstrap, simulated maximum likelihood or simulated method of moments. Even if a single execution does not take much time, repeating such a procedure for thousands of times will take a non-trivial duration.

Of course, optimizing code takes human time. It is important to balance human time and machine time.

Example

The heteroskedastic-robust variance for the OLS regression is

$$(X'X)^{-1}X'\hat{e}\hat{e}'X(X'X)^{-1}$$

The difficult part is $X'\hat{e}\hat{e}'X = \sum_{i=1}^{n} \hat{e}_i^2 x_i x_i'$. There are at least 4 mathematically equivalent ways to compute it.

1. literally sum over $i = 1, \ldots, n$ one by one.
2. $X'\text{diag}(\hat{e}^2)X$, with a dense central matrix.
3. $X'\text{diag}(\hat{e}^2)X$, with a sparse central matrix.
4. Do cross product to `X*e_hat`. This is different from the matrix formulation. It takes advantage of the element-by-element operation in R.

We first generate the data of binary response and regressors. Due to the discrete nature of the dependent variable, the error term in the linear probability model is heteroskedastic. It is necessary to use the heteroskedastic-robust variance to consistently estimate the asymptotic variance of the OLS estimator. The code chunk below estimates the coefficients and obtains the residual.

```
lpm = function(n){
    # estimation in a linear probability model

    # set the parameters
```

```
    b0 = matrix( c(-1,1), nrow = 2 )

    # generate the data
    e = rnorm(n)
    X = cbind( 1, rnorm(n) ) # you can try this line. See what is the difference.
    Y = (X %*% b0 + e >=0 )
    # note that in this regression b0 is not converge to b0 because the model is changed.

    # OLS estimation
    bhat = solve( t(X) %*% X, t(X)%*% Y )

    # calculate the t-value
    bhat2 = bhat[2] # parameter we want to test
    e_hat = Y - X %*% bhat
    return( list(X = X, e_hat = as.vector( e_hat) ) )
}
```

We run the 4 estimators for the same data, and compare the time.

```
# an example of robust matrix, sparse matrix. Vecteroization.
library(Matrix)
```

```
## Warning: package 'Matrix' was built under R version 3.3.2
```

```
# n = 5000; Rep = 10; # Matrix is quick, matrix is slow, adding is OK
n = 50; Rep = 1000;  # Matrix is slow,  matrix is quick, adding is OK

for (opt in 1:4){

  pts0 = Sys.time()

  for (iter in 1:Rep){

    set.seed(iter) # to make sure that the data used
    # different estimation methods are the same

    data.Xe = lpm(n)
    X = data.Xe$X;
    e_hat = data.Xe$e_hat

    XXe2 = matrix(0, nrow = 2, ncol = 2)

    if (opt == 1){
      for ( i in 1:n){
        XXe2 = XXe2 + e_hat[i]^2 * X[i,] %*% t(X[i,])
      }
    } else if (opt == 2) {# the vectorized version
      e_hat2_M = matrix(0, nrow = n, ncol = n)
      diag(e_hat2_M) = e_hat^2
      XXe2 = t(X) %*% e_hat2_M %*% X
    } else if (opt == 3)  {# the vectorized version
      e_hat2_M = Matrix( 0, ncol = n, nrow = n)
      diag(e_hat2_M) = e_hat^2
      XXe2 = t(X) %*% e_hat2_M %*% X
    } else if (opt == 4)  {# the best vectorization method. No waste
```

```
      Xe = X * e_hat
      XXe2 = t(Xe) %*% Xe
    }

    XX_inv = solve( t(X) %*% X  )
    sig_B =  XX_inv %*% XXe2 %*% XX_inv
  }
  cat("n = ", n, ", Rep = ", Rep, ", opt = ", opt, ", time = ", Sys.time() - pts0, "\n")
}
```

```
## n =  50 , Rep =  1000 , opt =  1 , time =  0.612432
## n =  50 , Rep =  1000 , opt =  2 , time =  0.1451139
## n =  50 , Rep =  1000 , opt =  3 , time =  2.100621
## n =  50 , Rep =  1000 , opt =  4 , time =  0.1120789
```

We clearly see the difference in running time, though the 4 methods are mathematically the same. When n is small, `matrix` is fast and `Matrix` is slow; the vectorized version is the fastest. When n is big, `matrix` is slow and `Matrix` is fast; the vectorized version is still the fastest.

Efficient Loop

In standard `for` loops, we have to do a lot of housekeeping work. `plyr`, developed by Hadley Wickham, simplifies the job and facilitates parallelization.

Example

Here we calculate the empirical coverage probability of a Poisson distribution of degrees of freedom 2. We first write a user-defined function `CI` for confidence interval, which was used in the last lecture.

```
CI = function(x){ # construct confidence interval
  # x is a vector of random variables

  n = length(x)
  mu = mean(x)
  sig = sd(x)
  upper = mu + 1.96/sqrt(n) * sig
  lower = mu - 1.96/sqrt(n) * sig
  return( list( lower = lower, upper = upper) )
}
```

This is a standard `for` loop.

```
Rep = 10000
sample_size = 1000

# a standard loop
out = rep(0, Rep)
pts0 = Sys.time() # check time
mu = 2
for (i in 1:Rep){
  x = rpois(sample_size, mu)
  bounds = CI(x)
  out[i] = ( ( bounds$lower <= mu  ) & (mu <= bounds$upper) )
}
cat( "empirical coverage probability = ", mean(out), "\n") # empirical size
```

```
## empirical coverage probability =  0.9522
```

```
pts1 = Sys.time() - pts0 # check time elapse
print(pts1)
```

```
## Time difference of 0.818573 secs
```

Pay attention to the line `out = rep(0, Rep)`. It *pre-defines* a vector `out` to be filled by `out[i] = ( ( bounds$lower <= mu  ) & (mu <= bounds$upper) )`. The computer opens a continuous patch of memory for the vector `out`. When new result comes in, the old element is replaced. On the contrary, If we do not pre-define `out` but append one more element in each loop, the length of `out` will change in each replication and every time a new patch of memory will assigned to store it. The latter approach will spend much more time just to locate the vector in the memory.

`out` is the result container. In a `for` loop, we pre-define a container, and replace the elements of the container in each loop by explicitly calling the index.

In contrast, a `plyr` loop saves the house keeping chores, and makes it easier to parallelize. In the example below, we encapsulate the chunk in the `for` loop as a new function `capture`, and run the replication via `__ply`. `__ply` is a family of functions. `ldply` here means that the input is a list (`l`) and the output is a data frame (`d`) .

```
library(plyr)

capture = function(i){
  x = rpois(sample_size, mu)
  bounds = CI(x)
  return( ( bounds$lower <= mu  ) & (mu <= bounds$upper) )
}

pts0 = Sys.time() # check time
out = ldply( .data = 1:Rep, .fun = capture )
cat( "empirical coverage probability = ", mean(out$V1), "\n") # empirical size
pts1 = Sys.time() - pts0 # check time elapse
print(pts1)
```

This example is so simple that the advantage of `plyr` is not be dramatic. The difference in coding will be noticeable in complex problems with big data frames. In terms of speed, `plyr` does not run faster than a `for` loop. They are of similar performance. Parallel computing will be our next topic. It is quite easy to implement parallel execution with `plyr`—we just need to change one argument in the function.

Parallel Computing

The packages `foreach` and `doParallel` are useful for parallel computing. Below is the basic structure. `registerDoParallel(number)` prepares a few CPU cores to accept incoming jobs. When calling `__ply`, we explicitly specify `.parallel` to be `TRUE`. If `myfunction` also depends on other variables in the session and/or other packages, we must export them to other cores via the argument `.paropts` (parameter options).

```
library(plyr)
library(foreach)
library(doParallel)

registerDoParallel() # opens other CPUs

l_ply(.data = 1:10,
      .fun = myfunction,
```

```
      .parallel = TRUE,
      .paropts = list( .packages = package.list,
        .export = ls(envir=globalenv() ) )
)
```

In theory, if we have two CPUs running simultaneously, we cut the time to a half of that on a single CPU. Is that what happening in practice?

Example

Compare the speed of a parallel loop and a single-core sequential loop.

```
registerDoParallel(2) # open 2 CPUs

pts0 = Sys.time() # check time
out = ldply(.data = 1:Rep, .fun = capture, .parallel = T,
            .paropts = list(.export = ls(envir=globalenv() )) )
cat( "empirical coverage probability = ", mean(out$V1), "\n") # empirical size
pts1 = Sys.time() - pts0 # check time elapse
print(pts1)
```

Surprisingly, the above code block of parallel computing runs even more slowly. It is because the task in each loop can be done in very short time. In contrast, the code chunk below will tell a different story. There the time in each loop is non-trivial, and then parallelism dominates the overhead of the CPU communication.

```
Rep = 200
sample_size = 2000000
pts0 = Sys.time() # check time
out = ldply(.data = 1:Rep, .fun = capture, .parallel = FALSE)
cat( "empirical coverage probability = ", mean(out$V1), "\n") # empirical size
pts1 = Sys.time() - pts0 # check time elapse
print(pts1)

# compare to the parallel version
pts0 = Sys.time() # check time
out = ldply(.data = 1:Rep, .fun = capture, .parallel = TRUE,
            .paropts = list(.export = ls(envir=globalenv() )) )
cat( "empirical coverage probability = ", mean(out$V1), "\n") # empirical size
pts1 = Sys.time() - pts0 # check time elapse
print(pts1)
```

Remote Computing

Investing money from our own pocket to a powerful laptop to conduct heavy-lifting computational work will surely be a regret. (i) We do not run these long jobs every day, it is more cost efficient to share a workhorse. (ii) We cannot keep our laptop always on when we move it around. The right solution is remote computing on a server.

No fundamental difference lies between local and remote computing. We prepare the data and code, open a shell for communication, run the code, and collect the results. One potential obstacle is dealing with a command-line-based operation system. Such command line tools is the norm of life two or three decades ago, but today we mostly work in a graphic operating system like Windows or OSX. For Windows users (I am one of them), I recommend PuTTY, a shell, and WinSCP, a graphic interface for input and output.

Here are a few commands for basic operating in a Unix/Linux operating system.

- mkdir

- cd
- copy
- top
- screen
- ssh user@address
- start a program

Our CUHK department runs two servers, one for students and the other for professors. The one for professors is a 16-core server. I have opened an account for you. You can try out this script on `econsuper`.

1. Log in `econsuper.econ.cuhk.edu.hk`;
2. Save the code block below as `loop_server.R`, and upload it to the server;
3. In a shell, run `R --vanilla <loop_server.R> result_your_name.out`;
4. To run a command in the background, add `&` at the end of the above command. To keep it running after closing the console, add `nohup` at the beginning of the command.

```
library(plyr)
library(foreach)
library(doParallel)

# prepare the functions
mu = 2
CI = function(x){
  # x is a vector of random variables

  n = length(x)
  mu = mean(x)
  sig = sd(x)
  upper = mu + 1.96/sqrt(n) * sig
  lower = mu - 1.96/sqrt(n) * sig
  return( list( lower = lower, upper = upper) )
}

capture = function(i){
  x = rpois(sample_size, mu)
  bounds = CI(x)
  return( ( bounds$lower <= mu  ) & (mu <= bounds$upper) )
}

############ implementation ###############
Rep = 200
sample_size = 5000000
pts0 = Sys.time() # check time
out = ldply(.data = 1:Rep, .fun = capture, .parallel = FALSE)
cat( "empirical coverage probability = ", mean(out$V1), "\n") # empirical size
pts1 = Sys.time() - pts0 # check time elapse
print(pts1)

# compare to the parallel version
registerDoParallel(16) # opens other CPUs

pts0 = Sys.time() # check time
out = ldply(.data = 1:Rep, .fun = capture, .parallel = TRUE,
            .paropts = list(.export = ls(envir=globalenv() )) )
cat( "empirical coverage probability = ", mean(out$V1), "\n") # empirical size
```

```
pts1 = Sys.time() - pts0 # check time elapse
print(pts1)
```