my-vignette

```
library(blblm)
```

###For more details on a spcific function, type ?blblm for example.

1) Vignette:

The first change I made to the base package was adding this vignette. I will outline any further changes that I make to the package here,along with an explanation of the implementation and purpose of each of the changes.

2) BLBLM Parallelization:

The Second way that the package blblm is modified is the addition of the ability to use parallelization to compute the linear model coefficients. This is done by using the function blblm_par(formula,data,m,B,ncores) as apposed to the single-core variant, blblm(formula,data,m,B). The user can specify the value of ncores to be between 1 and the maximum nunber of cores they have available on their computer. This function, blblm_par, uses the furrr package, and the functions future_map and plan in order to compute the linear model traits across each of the m groups the data is divided into, B times by bootstrapping. The tasks are split across the number of cores speicified for ncores.

Note: This function is not always faster than blblm, and is often slower than blblm when the data set or number of bootstraps selected is small.

Example: fit <- blblm_par(mpg ~ wt * hp, data = mtcars, m = 3, B = 100, 2) coef(fit)

3) Parallel Benchmarks:

The third improvement made to the base package blblm is the addition of a function to benchmark and compare the runtime and memory usage of both blblm and blblm_par. This is done through the function blblm_benchmarks(formula, data, m, B). This function takes in the same arguments as blblm, and it first acquires the number of cores that your computer has available to the user, and sets it equal to max_cores. It then runs blblm, the single-core variant and saves the runtime and memory usage. Then it calculates the runtime and memory usage for 2 through the maximimum number of cores that your computer has in order to see a comparison between single and parallel computing. Finally, it returns a dataframe containing the runtime and the memory used to compute the lm in both the single-core variant, and for the parallel with 2 through max_cores number of cores used.

Example: blblm_benchmarks(mpg ~ wt * hp, data = mtcars, m = 3, B = 100)

4) Logistic regression:

The Fourth improvement I made to the base package blblm is adding a function to allow the user to compute the coefficients and other characteristics of a dataseet using a logistic regression model. This is done by using the function blb_logreg(formula, data, m = 3, B = 5000). Note that in the formula, the response variable should be in the form 0/1 as the logistic regression model is a binary classifier. It uses a very similar structure to blblm in order to divide the data into m roughly equal sizes, bootstrap the data, and calculate the coefficients, sigma, etc. It is recommended to either use the default m = 3, or a small number for m, as this will keep the group sizes large, which will minimize the risk of the data being unseparble due to anomolies or perfectly separable, giving non reliable estimates.

Example: data <- iris labels <- rep(0:1,75) data$Species <- labels fit <- blb_logreg(Species ~ Petal.Length * Sepal.Length, data = data, m = 2, B = 100) coef(fit)

5) User-selected files:

The next improvement I mnade to the base package was adding the ability for the user to specify which files to use as the data. This is done by the function input_file, which takes in the pathname of the location where the data is found. The user performs this by specifying the location of folder containing csv files, and the function will read all of the csv files and use them as the data_files, rather than randomly splitting the data into m groups. To use this function, run blblm_par_user(formula,path,ncores, B), where the path is set to the directory of the folder containing the csv files.

Example: dir.create("files", showWarnings = FALSE) 1:100 %>% walk(function(i) { dt <- data.frame(x = rnorm(5000), y = rnorm(5000)) write_csv(dt, file.path("files", sprintf("file%02d.csv", i))) }) fit <- blblm_par_user(y ~ x, path = 'files',ncores = 2, B = 100) co <- coef(fit)

6) Documentation and Descriptions:

Added extensive documentation for each of the functions, inlcuding a name, a description, the input paramters, the return, and examples. This gives the user a more detailed view into each function that they are interested in.

Example: ?blblm

7) Tests:

Added tests for every function that I added to the package to verify that each work properly. I also added tests for several of the already existing functions.

Example: test-blblm.r

8) Check():

The package passes check() with one note pertaining to the directory used for the files used in the blm_par_use(user-selected file, parallel blblm), and the tests for all of the functions pass.