

blblm-vignette

Author: Xiaoqi Na

Email: xna@ucdavis.edu

Date: 06/10/2020

blblm

blblm & blbdm

This package is a implementation of the bag of little bootstrap for the linear regression models and logistic linear regression model. While the idea of bag of bootstrap for normal dataset is prevalent, we want to figure out how can it implement for a regression model inside the datasets. This process combines the usage of bootstrap and subsamples together, and it can give users a more efficient way to compute the estimators of the linear regression model. Moreover, this package also allows users to use parallelization for their model, and they can choose how many CPUs they want to use.

How does it work?

This is a common question for the blblm package. How does it work when we have a existing dataset? Firstly, split the dataset into m parts of approximated equal sizes, this procedure is called subsampling. Next, repeat the iteration of each subsamples B times by using `replicate()`. Then, we will get a blblm model from the procedures above. Eventually, we can compute the estimators of the the bag of little bootstrap for the linear regression models or logistic linear regression model, such as coefficients, sigma, confidence intervals and predictions for new data.

Set up the package

```
library(blblm)
```

Functions

There are two fitted functions (`blblm` and `blblm_parallel`) inside this packages to make the users to choose whether they want to use the parallization or not. Moreover, there are other functions for the estimators after fitting the `blblm` model, functions such as coefficient, sigma, confident interval and predictions.

Functions

`blblm()` & `blbglm()`

This function is a normal function to get the fit of bag of little bootstrap for the linear regression model.

- `formula`: The formula of the linear regression model or logistic linear regression model. It usually has a form `response ~ terms`, where `response` is usually one numeric vector or one column from a dataset, and `term` is a series of terms which might have a linear relationship with the response.
- `data`: The dataset or a dataframe that users want to analyze.
- `parallel`: A logistic Value (TRUE/FALSE). Whether users want to use parallelization to use more than one CPUs. If you choose to use parallization, run `plan(multisession, worker=i)` first, where `i` is the number of CPUs you are willing to use.
- `m`: It is a numeric, and it splits the dataset into `m` parts of approximated equal sizes, this procedure is subsampling.
- `B`: It is a numeric, and it repeats the subsampling process `B` times. `*...:` Other arguments

Estimators' Functions

`print.blblm()` & `print.blbglm`

This function can print out the `blblm` model the user want to get.

- `x`: the fit of the `blblm` or `blbglm` model that generated from the above `blblm` or `blbglm` function. `*...:` Other arguments

`coef.blblm()` & `coef.blbglm()`

This function is to compute the coefficients of the blblm model.

- object: the fit of the model that generated from the above `blblm` or `blbglm` function. *...: Other arguments

sigma.blblm() & sigma.blbglm()

This function is to compute the sigma of the blblm model. Also, the user can choose whether use the confidence interval or not for the sigma estimator.

- object: the fit of the blblm model that generated from the above `blblm` or `blbglm` function.
- confidence: It is a logical, users can choose whether they want the output has confidence intervals of sigma.
- level: If the user choose to have confidence interval in their output, it is the level of the confidence they user want to choose. *...: Other arguments

confint.blblm() & confint.blbglm

This function it to compute the bag of little bootstrap confidence interval for each term in the linear regression model.

- object: the fit of the blblm model that generated from the above `blblm` or `blbglm` function.
- parm: It is a specification of which parameters are to be given confidence intervals. Also, the user can choose NULL for this parm, which means all parameters will be taken into consideration.
- level: the level of confidence. Eg.0.95 *...: Other arguments

predict.blblm() & predict.blbglm()

This function it to compute the prediction based on the bag of little bootstraps for linear regression models

- object: the fit of the blblm or blbglm model
- new_data: It is the another data frame that the users want to look for the predictions.
- confidence: It is a logical, users can choose whether they want the output has confidence intervals of sigma.
- level: If the user choose to have confidence interval in their output, it is the level of the confidence they user want to choose. *...: Other arguments

Exmaples

Linear Regression Model Example:

```
fit <-  
  blblm(formula=Sepal.Length~Petal.Length+Petal.Width,data=iris,m=10,B=100,parallel=TRUE)  
summary(fit)
```

```
#>           Length Class  Mode
#> estimates 10      -none- list
#> formula    3      formula call
```

```
coef.blblm(fit)
```

```
sigma.blblm(fit, confidence=TRUE, level=0.95)
```

```
confint.blblm(fit, c("Petal.Length", "Petal.Width"), level = 0.95, parm=NULL)
```

```
predict.blblm(fit,
```

```
data.frame(Petal.Length=c(3.0, 5.0), Petal.Width=c(0.5, 2.0)), confidence=TRUE, level=0.95)
```

Logistic Linear Regression Model Example:

```
fitglm <-
  blbgglm(formula=Sepal.Length~Petal.Length+Petal.Width, data=iris, m=10, B=100, parallel=TRUE)
summary(fitglm)
#>           Length Class  Mode
#> estimates 10      -none- list
#> formula    3      formula call
```

```
coef.blbgglm(fit)
```

```
sigma.blbgglm(fit, confidence=TRUE, level=0.95)
```

```
confint.blbgglm(fit, c("Petal.Length", "Petal.Width"), level = 0.95, parm=NULL)
```

```
predict.blbgglm(fit,
```

```
data.frame(Petal.Length=c(3.0, 5.0), Petal.Width=c(0.5, 2.0)), confidence=TRUE, level=0.95)
```

Ancillary functions inside the packages

split_data()

This function is to subsample the dataset by splitting the dataset into m parts of approximated equal sizes

lm1() & glm1()

This function is to estimate the regression estimates based on given the number of repetitions, and it will be used in the `lm_each_boot()` or `glm_each_boot()` function.

lm_each_boot() & glm_each_boot()

This function is to compute the regression estimates for a blb dataset, and this will be further used in the `lm_each_subsample` function.

lm_each_subsample() & glm_each_subsample()

This function is to compute the estimates, and repeat it **B** times. And, this function is useful in `blblm()` or `blbglm` function.

There are also some other helpful functions computing the mean, lower and upper bound using for this blblm package.

```
mean_lwr_upr() map_mean map_cbind map_rbind
```

Additional informations

If users want to know more about the functions inside the packages, they can do

```
?blblm
```

```
?blblm_parallel
```

```
?coef.blblm
```

```
?sigma.blblm
```

```
?confint.blblm
```

```
?predict.blblm ### Also, similar for the logistic regression model, just change blblm to blbglm.
```

Reference

<https://github.com/ucdavis-sta141c-sq-2020/sta141c-notes/tree/master/ch08-bootstrap>

`lm` function in `r`
