Computational Statistics Assignment

Bootstrapping

Luycer Bosire

4/14/2021

#1. Explain what bootsrapping is.

The bootstrap is a widely applicable and extremely powerful statistical tool used to quantify the uncertainty associated with a given estimate or statistical learning. The bootstrap resampling method can be used to measure the accuracy of a predictive model. It can be used to estimate the standard errors of the coefficient from a linear regression fit.

The bootstrap algorithm begins by generating a large number of independent bootstrap samples each of size Typical values for the number of bootstrap samples, range from to for standard error estimation.

Corresponding to each bootstrap sample is a *bootstrap replication* of namely the value of the statistic evaluated for

If is the sample median, for instance, then is the median of the bootstrap sample. The bootstrap estimate of standard error is the standard deviation of the bootstrap replications,

where

#2. Use examples to show the difference between parametric and non-parametric bootsrap.

Non-parametric bootstrapping is where the data under analysis comes from the same distribution and sample size as the data at hand and thus is sampled with replacement to preserve the probability density function from the dataset itself.

A brief overview of the non-parametric bootstrap method below serves as our example;

Given

) iid from

.

is an estimator of

and and

is a centered or standardized random variable constructed from for example;

where

is a known standard deviation or estimated standard deviation, say

.

Consider

iid from

, the empirical distribution function of

. Based on these iid random variables construct using the same formula as for

and

that is the same as ;

and

Parametric bootsrapping assumes that the data comes from a known distribution with unknown parameters. The parameters are estimated from the data at hand and used to estimate distributions to simulate the samples.

A brief overview of the parametric bootstrap method below serves as our example;

Given

) iid from

.

is used to denote the true value of the parameter.

is an estimator of

and

is a centered or standardized random variable constructed from for example;

where

is a known standard deviation or estimated standard deviation, say

.

Consider

iid from

. Based on these iid random variables construct using the same formula as for

and

that is the same as ;

and

The difference between parametric and non-parametric lies in the distribution of the sample

We use Monte Carlo simulation with **M** simulation steps of size *n* each to approximate the sampling distribution of

for this given

.

The parametric bootstrap theory is that

and

have approximately the same distribution, thus approximately the same quantiles.

#3. Explain the bootsrap Confidence Interval.

Using an example of a population from a normal distribution of sample size

and parametric boostraping resampling;

1. Compute the sample mean
2. and variance
3. . The bootsrap samples can be taken by generating random samples of size
4. from
5. .
6. Suppose we take a sample of 1000, the set of 1000 bootstrap sample means should be a good estimate of the sampling distribution of x.
7. A 95% CI for the population mean is then formed by sorting the bootstrap means from lowest to highest, and dropping the 2.5% smallest and largest remaining values at the ends of the CI.

Based on the example in *question 2* we can obtain quantiles for the central 0.95 region by solving for

from

#4. Use the in-built function boot to illustrate various aspects of boostrapping.

Performing a bootstrap analysis in R entails two steps. First, we create a function that computes the statistic of interest. Then we use the **boot()** function, which is be of the **boot** library to perform the bootstrap by repeatedly sampling observations from the dataset with replacement.

We make use of the **Portfolio** dataset in the package **ISLR**.

First we create a function **bootanalysis** that takes input X and Y data and a vendor indicating which observations should be used to estimate a parameter

. The function then outputs the estimate for

based on the selected observations.

library(ISLR)

## Warning: package 'ISLR' was built under R version 4.0.5

attach(Portfolio)  
  
bootanalysis<- function(data,index){  
X<- data$X[index]  
Y<- data$Y[index]  
return((var(Y)-cov(X,Y))/(var(X)+var(Y)-2\*cov(X,Y)))}

This function returns an estimate for

. For example, the following command tells

to estimate

using all 100 observations.

set.seed(1)  
bootanalysis(Portfolio, sample(100,100, replace = T))

## [1] 0.7368375

The above procedure will require us to perform the command many times, recording all of the corresponding estimates for

and computing the resulting standard deviation.

We produce R=1000 bootstrap estimates for

library(boot)  
boot(Portfolio, bootanalysis, R=1000)

##   
## ORDINARY NONPARAMETRIC BOOTSTRAP  
##   
##   
## Call:  
## boot(data = Portfolio, statistic = bootanalysis, R = 1000)  
##   
##   
## Bootstrap Statistics :  
## original bias std. error  
## t1\* 0.5758321 -0.001695873 0.09366347

The output shows that using the original data,

and that the bootstrap estimate for

Next we analyze the accuracy of a linear regression model. We assess the variability of the estimates for

and

, the intercept and the slope terms for the linear model

bootanalysis<- function(data, index){  
 return(coef(lm(mpg~horsepower, data = data, subset = index)))  
}  
  
bootanalysis(Auto, 1:392)

## (Intercept) horsepower   
## 39.9358610 -0.1578447

Next we create bootstrap estimates for the intercept and slope terms by randomly sampling from among the observations with replacement.

set.seed(1)  
bootanalysis(Auto, sample(392,392, replace = T))

## (Intercept) horsepower   
## 40.3404517 -0.1634868

bootanalysis(Auto, sample(392,392, replace = T))

## (Intercept) horsepower   
## 40.1186906 -0.1577063

We then use the *boot()* function to compute the standard errors of 1000 bootstrap estimates for the intercept and slope terms

boot(Auto, bootanalysis, 1000)

##   
## ORDINARY NONPARAMETRIC BOOTSTRAP  
##   
##   
## Call:  
## boot(data = Auto, statistic = bootanalysis, R = 1000)  
##   
##   
## Bootstrap Statistics :  
## original bias std. error  
## t1\* 39.9358610 0.0544513229 0.841289790  
## t2\* -0.1578447 -0.0006170901 0.007343073

This indicates that the bootstrap estimate for

#5. Using wages dataset in R and use the bootstrap estimates to estimate the variance of the sample median as well as the 95% CI for the population median.

library(boot)  
library(ISLR)  
  
data(Wage)  
median(Wage$wage)

## [1] 104.9215

var(Wage$wage)

## [1] 1741.276

data\_medians<- function(x, indices){  
 return(median(x[indices]))  
}  
wage\_bootstrapanalysis<- boot(Wage$wage, data\_medians, 10000)  
  
#We compute the original sample median  
median(Wage$wage, 1:length(Wage$wage))

## Warning in if (na.rm) x <- x[!is.na(x)] else if (any(is.na(x))) return(x[FALSE]  
## [NA]): the condition has length > 1 and only the first element will be used

## [1] 104.9215

wage\_bootstrapanalysis$t0

## [1] 104.9215

#Note that the 2 medians are similar  
  
wage\_bootstrapanalysis$R

## [1] 10000

wage\_bootstrapanalysis$call

## boot(data = Wage$wage, statistic = data\_medians, R = 10000)

boot(data = Wage$wage, statistic = data\_medians, R = 10000)

##   
## ORDINARY NONPARAMETRIC BOOTSTRAP  
##   
##   
## Call:  
## boot(data = Wage$wage, statistic = data\_medians, R = 10000)  
##   
##   
## Bootstrap Statistics :  
## original bias std. error  
## t1\* 104.9215 0.284396 0.6626873

boot.ci(wage\_bootstrapanalysis)

## Warning in boot.ci(wage\_bootstrapanalysis): bootstrap variances needed for  
## studentized intervals

## Warning in norm.inter(t, adj.alpha): extreme order statistics used as endpoints

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS  
## Based on 10000 bootstrap replicates  
##   
## CALL :   
## boot.ci(boot.out = wage\_bootstrapanalysis)  
##   
## Intervals :   
## Level Normal Basic   
## 95% (103.3, 106.0 ) (102.9, 105.8 )   
##   
## Level Percentile BCa   
## 95% (104.0, 106.9 ) (101.8, 104.9 )   
## Calculations and Intervals on Original Scale  
## Warning : BCa Intervals used Extreme Quantiles  
## Some BCa intervals may be unstable