Bootstrapping in R – A Tutorial

Eric B. Putman

Department of Ecosystem Science and Management
Bootstrapping

• Resampling technique with replacement
  – “The population is to the sample as the sample is to the bootstrap samples”

• Allows estimation of the sampling distribution of a statistic
  – Confidence intervals, bias, variance, etc.
Procedure

- Resample a dataset a given number of times
- Calculate a statistic from each sample
- Accumulate the results and calculate the sample distribution of the statistic
Objective

• Calculate a series of linear regressions to determine which variable or combination of variables best explains the volume of black cherry trees
  – Comparisons made using coefficient of determination (R-squared)

• Bootstrap the linear regressions (for each bootstrap sample) to determine 95% confidence intervals of their respective R-squared values
Data

• “trees” dataset (included in R)
• Volume (cubic feet), girth (diameter in inches, measured at breast height), and height (feet) measurements of 31 felled black cherry trees

help(trees)
Code Walkthrough
• Load the boot library
  – Contains functions to conduct bootstrapping

```r
library(boot)
```

• Investigate the “trees” dataset

```r
head(trees)
```

```
   Girth Height  Volume
 1    8.3     70  10.3
 2    8.6     65  10.3
 3    8.8     63  10.2
 4   10.5     72  16.4
 5   10.7     81  18.8
 6   10.8     83  19.7
```
• Explore relationships between volume, girth, and height

\[
\text{plot(trees$Volume \sim trees$Height, main = 'Black Cherry Tree Volume Relationship', xlab = 'Height', ylab = 'Volume', pch = 16, col = 'blue')}
\]

\[
\text{plot(trees$Volume \sim trees$Girth, main = 'Black Cherry Tree Volume Relationship', xlab = 'Girth', ylab = 'Volume', pch = 16, col = 'blue')}
\]
• Create a function that will calculate a statistic (or multiple statistics) on each bootstrap sample

• Function syntax in R

```r
foo = function(parameter_1, parameter_2,... parameter_n){
    bar = *do something to data passed as parameters*
    return(bar)
}
```
Statistic-calculation function for the boot package takes two specific parameters (simple example) and will be applied to each bootstrap sample

```r
calculate = function(data, indices) {
    sample = data[indices, ]
    bar = mean(sample)
    return(bar)
}
```

Creates the bootstrap sample (i.e., subset the provided data by the "indices" parameter). "indices" is automatically provided by the "boot" function; this is the sampling with replacement portion of bootstrapping

```r
calculate = function(data, indices) {
    return(mean(data[indices]))
}
```

Or, more concisely:
• Create a function to calculate linear regressions of several variable combinations and return their respective R-squared values
  – Height only,
  – Girth only
  – Girth / height ratio
  – Girth and height
  – Girth, height, and girth / height ratio

• Note that we are calculating (and returning) multiple statistics simultaneously
  – These statistics will be calculated for each bootstrap sample

```r
volume_estimate = function(data, indices){
  d = data[indices, ]
  H_relationship = lm(d$Volume~d$Height, data = d)
  H_r_sq = summary(H_relationship)$r.square
  G_relationship = lm(d$Volume~d$Girth, data = d)
  G_r_sq = summary(G_relationship)$r.square
  G_H_ratio = d$Girth / d$Height
  G_H_relationship = lm(d$Volume~G_H_ratio, data = d)
  G_H_r_sq = summary(G_H_relationship)$r.square
  combined_relationship = lm(d$Volume~d$Height + d$Girth, data = d)
  combined_r_sq = summary(combined_relationship)$r.square
  combined_2_relationship = lm(d$Volume~d$Height +d$Girth + G_H_ratio, data = d)
  combined_2_r_sq = summary(combined_2_relationship)$r.square
  relationships = c(H_r_sq, G_r_sq, G_H_r_sq, combined_r_sq, combined_2_r_sq)
  return(relationships)
}
```
Statistics are added to a vector, which is then returned to the “boot” function
• Conduct the bootstrapping
  – Use “boot” function

```r
results = boot(data = trees, statistic = volume_estimate, R = 5000)
```

**Dataset from which statistics will be calculated**

**Function we created to calculate statistics on each bootstrap sample**

**Number of bootstrap samples (i.e., iterations)**
• View some calculated statistics of boot object

```r
print(results)
```

```
ORDINARY NONPARAMETRIC BOOTSTRAP

Call:
boot(data = trees, statistic = volume_estimate, R = 5000)

Bootstrap Statistics :
                 original     bias   std. error
  t1*  0.3579026  0.0024051943  0.12025420
  t2*  0.9353199  0.0005495767  0.01751679
  t3*  0.7309204  0.0025156062  0.08064029
  t4*  0.9479500  0.0032851681  0.01210484
  t5*  0.9732894  0.0005447157  0.01042662
```

$t^*$ corresponds to index of “relationships” vector (e.g., $t1^*$ refers to height only R-squared value
• Plot the boot objects
  – Provides histogram and Q-Q plot

```r
plot(results, index = 1)
```

The index parameter corresponds to the indices of the vector ("relationships") returned by the "volume_estimation" function (e.g., index 1 is the first item in the vector, which is the height only R-squared value)

```r
relationships = c(H_r_sq, G_r_sq, G_H_r_sq, combined_r_sq, combined_2_r_sq)
```

Height only R-squared distribution:
• Calculate 95% confidence intervals for each of the bootstrapped R-squared values
  – Using “Bias Corrected and Accelerated” (BCa) method

Specify index corresponding to position in vector for each statistic

```
confidence_interval_H = boot.ci(results, index = 1, conf = 0.95, type = 'bca')
print(confidence_interval_H)
ci_H = confidence_interval_H$bca[, c(4, 5)]
print(ci_H)
```

Store confidence intervals in a variable in order to plot later

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 5000 bootstrap replicates
CALL : boot.ci(boot.out = results, conf = 0.95, type = "bca", index = 1)

Intervals :
Level     BCa
95%   ( 0.1415,  0.6123 )
Calculations and Intervals on Original Scale
> ci_H = confidence_interval_H$bca[, c(4, 5)]
> print(ci_H)
0.1414861 0.6122950
```
• View histograms (frequency and density)
• Add kernel density line (blue)
• Add 95% confidence intervals (red)

```r
hist(results$t[,1], main = 'Coefficient of Determination: Height', xlab = 'R-Squared', col = 'grey')
hist(results$t[,1], main = 'Coefficient of Determination: Height', xlab = 'R-Squared', col = 'grey', prob = T)
lines(density(results$t[,1]), col = 'grey', prob = T)
abline(v = ci_H, col = 'red')
```

Note syntax to call desired sample distribution
• Can also call the entire sample distribution to further manipulate, save, etc.

```r
can also call the entire sample distribution to further manipulate, save, etc.
```

Access the sample statistics of each bootstrap sample

```r
results$t[, 1]
```

Subset to particular statistic; first column of the boot object "t" corresponds to the first item in the vector returned by the "volume_estimate" function

R-squared values of height only linear regression:

```
> results$t[, 1]
 [1] 0.207990443 0.363816239 0.579971818 0.423443272 0.336572704 0.417656521 0.251820295 0.343777274
 [9] 0.270477273 0.480302587 0.564330760 0.474092665 0.174531538 0.300817972 0.502245182 0.359519760
[17] 0.367795668 0.435299147 0.243218209 0.180413913 0.428146329 0.568726861 0.399806911 0.195195281
[25] 0.258577036 0.416366115 0.315921685 0.54119595 0.272757355 0.628962441 0.350397269 0.192770891
[33] 0.266364939 0.310743438 0.613576574 0.696147632 0.488130237 0.388040468 0.344063541 0.399933017
[41] 0.255363943 0.395594597 0.318028661 0.391665068 0.35077907 0.18840159 0.421280357 0.072206043
[49] 0.449664202 0.462657862 0.413759773 0.446951604 0.369800075 0.468153637 0.182068140 0.375718017
[57] 0.151727603 0.237096695 0.293074927 0.476329686 0.308111480 0.218648993 0.265019573 0.204667380
[65] 0.651896672 0.639127085 0.478180644 0.315661237 0.630257581 0.426617868 0.352848563 0.333865284
```
Results

• Linear regression with explanatory variables of girth, height, and girth / height ratio provided best coefficients of determination to model the volume of black cherry trees

• 5,000 sample bootstrap allowed estimation of R-squared sampling distribution
  – Could have also bootstrapped values of coefficients, additional models, etc.

| Estimating Black Cherry Tree Volume - Linear Regression Coefficients of Determination |
|-------------------------------------|---------|--------|-----------------|------------------|
| Original Value | Bias     | Std. Error | 95% Confidence Interval |
| Height Only     | 0.3579026 | 0.002405194 | 0.1202542 | 0.1414861 - 0.6122950 |
| Girth Only      | 0.9353199 | 0.000549577 | 0.01751679 | 0.8770796 - 0.9582597 |
| Girth / Height  | 0.7309204 | 0.002515606 | 0.08064029 | 0.4782823 - 0.8421099 |
| Girth and Height| 0.94795   | 0.003285168 | 0.01210484 | 0.9052392 - 0.9647783 |
| Girth, Height, and Girth / Height | 0.9732894 | 0.000544716 | 0.01042662 | 0.9418756 - 0.9868528 |
References

http://www.statmethods.net/advstats/bootstrapping.html
http://www.r-bloggers.com/bootstrap-example/
http://cran.r-project.org/web/packages/boot/boot.pdf