Ch 5.2: The Bootstrap

Lecture 16 - CMSE 381

Prof. Elizabeth Munch

Michigan State University
::
Dept of Computational Mathematics, Science & Engineering

Wed, Oct 12, 2023

Announcements

Last time:

- k-fold CV for Classification

Announcements:

- Fifth homework posted

Covered in this lecture

- Bootstrap

Section 1

The Bootstrap

The Idea

The goal: quantify the uncertainty associated with a given estimator or statistical learning method.

The boostrap idiom

"Pull yourself up by your bootstraps."

- Originally used as a saying meaning an impossible task, used sarcastically
- Now often used to imply that socioeconomic advancement is something everyone should be able to do

Today's class: Bootstrap on a simple modeling problem

- We wish to invest a fixed sum of money in two financial assets that yield returns of X and Y, respectively, where X and Y are random quantities.
- We will invest a fraction α of our money in X, and will invest the remaining $1 - \alpha$ in Y.
- Since there is variability associated with the returns on these two assets, we wish to choose α to minimize the total risk, or variance, of our investment.

One can show.....

...that $\mathrm{Var}(\alpha X + (1-\alpha)Y)$
is minimzed by

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$$

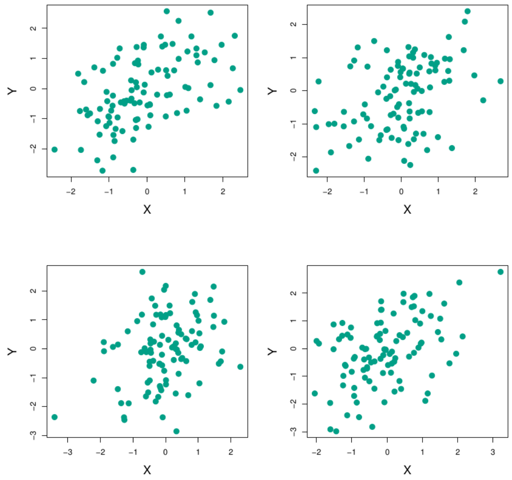
where

- $\sigma_X^2 = \mathrm{Var}(X)$
- $\sigma_Y^2 = \mathrm{Var}(Y)$
- $\sigma_{XY} = \mathrm{Cov}(X, Y)$

Get an estimate:

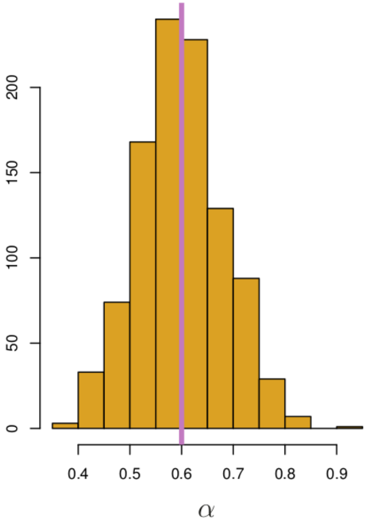
$$\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}}$$

Simulated data



- Simulate data using
  - $\sigma_X^2 = 1$
  - $\sigma_Y^2 = 1.25$
  - $\sigma_{XY} = 0.5$
  - Implies $\alpha = 0.6$
- In each panel: Simulate 100 pairs of returns for investments
- Predict σ_X^2, σ_Y^2, and σ_{XY}
- $\hat{\alpha}$ prediction by panel:
  0.576 0.543
  0.657 0.651

Resimulate: Rinse and repeat

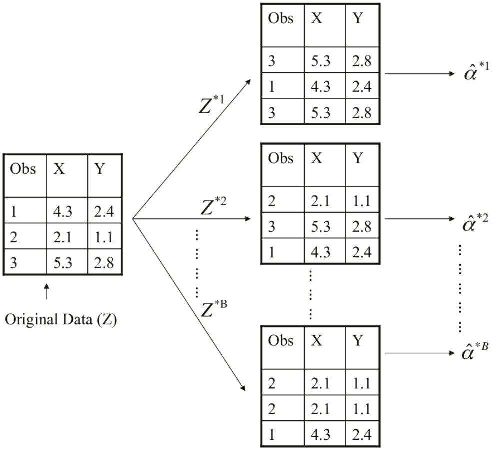


- Simulate 100 data points 1,000 times
- Left: Histogram of predictions for α
- Pink line: True value for α
- Mean over simulated values:
  $0.5996 = \bar{\alpha} = \frac{1}{1000} \sum \hat{\alpha}_r$
- St dev: $0.083 = \sqrt{\frac{1}{1000-1} \sum (\hat{\alpha}_r - \bar{\alpha})}$

Coding part 1: Apprximate $\hat{\alpha}$ with simulated data

So what's the problem?

The solution: Sample the data with replacement



Computation of error

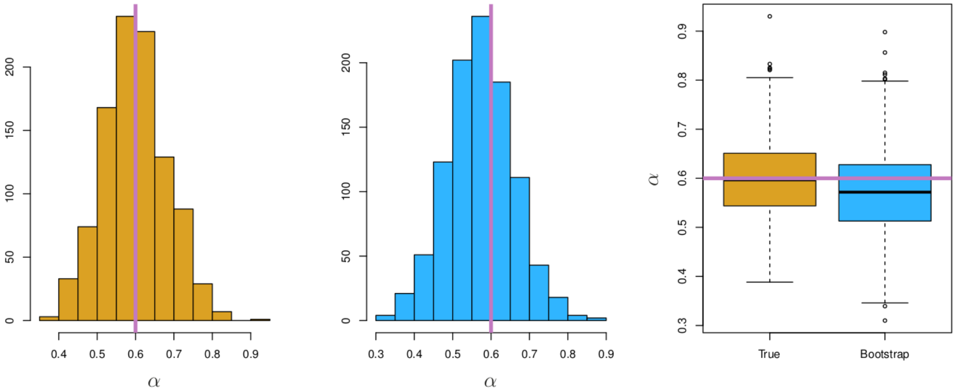
- Repeate procedure B times:
- Get B bootstrap data sets, $Z^{*1}, Z^{*2}, \ldots, Z^{*B}$
- Get B bootstrap estimates $\hat{\alpha}^{*1}, \hat{\alpha}^{*2}, \ldots, \hat{\alpha}^{*B}$

Get standard error estimate:

$$SE_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}\left(\hat{\alpha}^{*r} - \frac{1}{B}\sum_{r'=1}^{B}\hat{\alpha}^{*r'}\right)^2}$$

Coding part 2: Resampling data

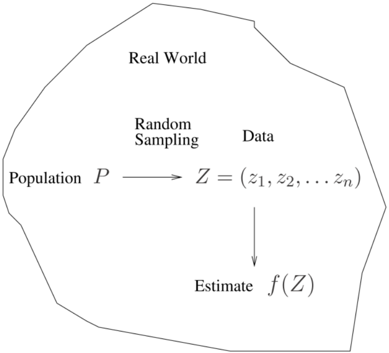
Back to the example

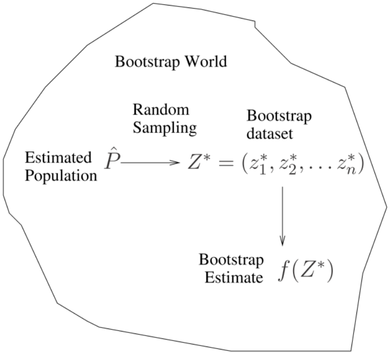


Resample version
Predicted $SE(\hat{\alpha}) = 0.083$

Bootstrap version
Predicted $SE(\hat{\alpha}) = 0.087$

A general picture for the bootstrap





TL;DR

- Start with data set of n points
- Sample n points with replacement to get data set Z^{*1}
- Use this to estimate whatever parameter we want $\hat{T}^{*1}$
- Repeat B times to get estimates $\hat{T}^{*1}, \cdots, \hat{T}^{*B}$
- Estimate standard error of our T estimate by

$$SE_B(\hat{T}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}\left(\hat{T}^{*r} - \frac{1}{B}\sum_{r'=1}^{B}\hat{T}^{*r'}\right)^2}$$

Boostrap vs Cross-Validation

Bootstrap:

CV:

Next time

12	Mon	Oct 2	Leave one out CV	5.1.1, 5.1.2	
13	Wed	Oct 4	k-fold CV	5.1.3	
14	Fri	Oct 6	More k-fold CV,	5.1.4-5	
15	Mon	Oct 9	k-fold CV for classification	5.1.5	HW #4 Due
16	Wed	Oct 11	Resampling methods: Bootstrap	5.2	
17	Fri	Oct 13	Subset selection	6.1	
18	Mon	Oct 16	Shrinkage: Ridge	6.2.1	
19	Wed	Oct 18	Shrinkage: Lasso	6.2.2	
	Fri	Oct 20	**Review**		
	Mon	Oct 23	No class - Fall break		
	Wed	Oct 25	**Midterm #2**		
20	Fri	Oct 27	Dimension Reduction	6.3	