ECON3389 Machine Learning in Economics

Module 3
Resampling Methods

Alberto Cappello

Department of Economics, Boston College

Fall 2024

Overview

Agenda:

- Bootstrap

Readings:

- ISLR Chapter 5

The Bootstrap

- The *bootstrap* is a flexible and powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning method when direct (exact) methods of inference are unavailable.
- For example, it can provide an estimate of the standard error of a coefficient in a regression model known to violate usual OLS assumptions required for exact or asymptotic inference.

The Bootstrap

- The *bootstrap* is a flexible and powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning method when direct (exact) methods of inference are unavailable.
- For example, it can provide an estimate of the standard error of a coefficient in a regression model known to violate usual OLS assumptions required for exact or asymptotic inference.
- The use of the term bootstrap derives from the phrase *to pull oneself up by one's bootstraps*, widely thought to be based on one of the eighteenth century "The Surprising Adventures of Baron Munchausen" by Rudolph Erich Raspe:

  The Baron had fallen to the bottom of a deep lake. Just when it looked like all was lost, he thought to pick himself up by his own bootstraps.

Bootstrap example

- Suppose that we wish to invest a fixed sum of money in two financial assets that yield random returns of X and Y.
- We will invest a fraction α of our money in X, and the remaining $(1 - \alpha)$ in Y.
- We wish to choose α to minimize the total risk, or variance, of our investment. In other words, we want to minimize $Var(\alpha X + (1 - \alpha)Y)$.

Bootstrap example

- Suppose that we wish to invest a fixed sum of money in two financial assets that yield random returns of X and Y.
- We will invest a fraction α of our money in X, and the remaining $(1-\alpha)$ in Y.
- We wish to choose α to minimize the total risk, or variance, of our investment. In other words, we want to minimize $Var(\alpha X + (1-\alpha)Y)$.
- One can show that the value that minimizes the risk is given by

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

where $\sigma_Y^2 = Var(Y)$, $\sigma_X^2 = Var(X)$ and $\sigma_{XY} = Cov(X, Y)$.

Bootstrap example

- But we do not know true population values of σ_X^2, σ_Y^2 and σ_{XY}.
- We can use a sample with measurements on X and Y to calculate estimated values of $\widehat{\sigma}_X^2$, $\widehat{\sigma}_Y^2$ and $\widehat{\sigma}_{XY}$.
- We can then estimate the optimal value of α that minimizes the variance of our investment as

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

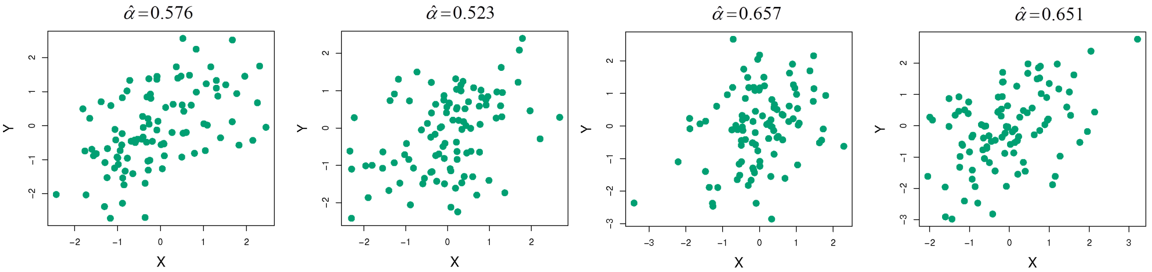
Bootstrap example

- But we do not know true population values of σ_X^2, σ_Y^2 and σ_{XY}.
- We can use a sample with measurements on X and Y to calculate estimated values of $\widehat{\sigma}_X^2$, $\widehat{\sigma}_Y^2$ and $\widehat{\sigma}_{XY}$.
- We can then estimate the optimal value of α that minimizes the variance of our investment as

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

- Even though there are know statistical inference results on exact distributions of $\widehat{\sigma}_X^2$, $\widehat{\sigma}_Y^2$ and $\widehat{\sigma}_{XY}$, it is hard or maybe even impossible to derive/proof an exact distribution of $\widehat{\alpha}$.
- We can, however, look at what values of $\widehat{\alpha}$ we can get if we simulate new samples of X and Y.

Bootstrap example



Each panel displays a separate sample of 100 simulated returns for investments X and Y, with corresponding $\widehat{\alpha}$ values displayed above each panel.

Bootstrap example

- Repeat the process of simulating 100 pairs of X and Y 1000 times, using the values of $\sigma_X^2 = 1$, $\sigma_Y^2 = 1.25$ and $\sigma_{XY} = 0.5$.
  - These values mean that the true value of α is 0.6

Bootstrap example

- Repeat the process of simulating 100 pairs of X and Y 1000 times, using the values of $\sigma_X^2 = 1$, $\sigma_Y^2 = 1.25$ and $\sigma_{XY} = 0.5$.
  - These values mean that the true value of α is 0.6
- Calculate 1000 corresponding values of $\widehat{\alpha}_1, \widehat{\alpha}_2, \ldots, \widehat{\alpha}_{1000}$. The resulting sample of 1000 estimates of $\widehat{\alpha}$ gives us the following mean and standard deviation:

$$\bar{\hat{\alpha}} = \frac{1}{1000} \sum_{r=1}^{1000} \widehat{\alpha}_r = 0.5996$$

$$SD(\hat{\alpha}) = \sqrt{\frac{1}{1000-1} \sum_{r=1}^{1000} \left(\widehat{\alpha}_r - \bar{\hat{\alpha}}\right)^2} \approx 0.083$$

- The simulations' average 0.5996 is very close to the true value of 0.6, with approximate 90% confidence interval of $(0.43; 0.74)$.

Bootstrap with real data

- The procedure outlined above cannot be applied to real life data, because for real data we cannot generate new samples from the original population.
- However, the bootstrap approach allows us to mimic the process of obtaining new data sets, so that we can quantify the variability of our estimates without generating additional new samples.

Bootstrap with real data

- The procedure outlined above cannot be applied to real life data, because for real data we cannot generate new samples from the original population.
- However, the bootstrap approach allows us to mimic the process of obtaining new data sets, so that we can quantify the variability of our estimates without generating additional new samples.
- Rather than repeatedly obtaining independent datasets from the population, we instead obtain distinct datasets by repeatedly sampling observations from the original data with replacement.
- Each of these bootstrap datasets is created by sampling *with replacement*, and is the *same size* as our original dataset.
- As a result some observations may appear more than once in a given bootstrap data set and some not at all.

Bootstrap example with a sample of size 3

- Our original sample data Z includes 3 pairs of X and Y.

Obs	X	Y
1	4.3	2.4
2	2.1	1.1
3	5.3	2.8

↑

Original Data (Z)

Bootstrap example with a sample of size 3

- Our original sample data Z includes 3 pairs of X and Y.
- We use the 1st bootstrap dataset Z^{*1} to produce a new bootstrap estimate $\widehat{\alpha}^{*1}$.

Obs	X	Y
1	4.3	2.4
2	2.1	1.1
3	5.3	2.8

Original Data (Z)

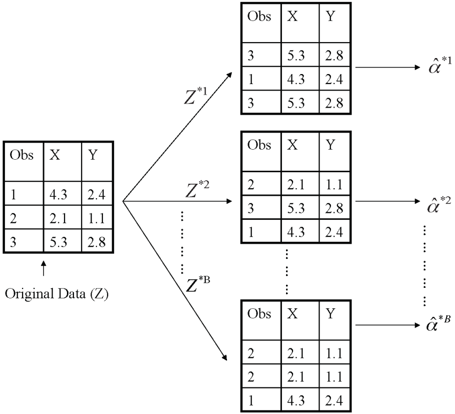
Z^{*1}

Obs	X	Y
3	5.3	2.8
1	4.3	2.4
3	5.3	2.8

$\hat{\alpha}^{*1}$

Bootstrap example with a sample of size 3

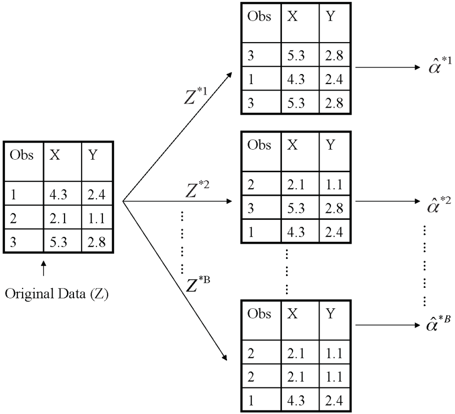
- Our original sample data Z includes 3 pairs of X and Y.
- We use the 1st bootstrap dataset Z^{*1} to produce a new bootstrap estimate $\widehat{\alpha}^{*1}$.
- This procedure is repeated B times (say, $B = 1000$) to produce B different bootstrap datasets $Z^{*1}, Z^{*2}, \dots, Z^{*B}$ and B corresponding estimates $\widehat{\alpha}^{*1}, \widehat{\alpha}^{*2}, \dots, \widehat{\alpha}^{*B}$.

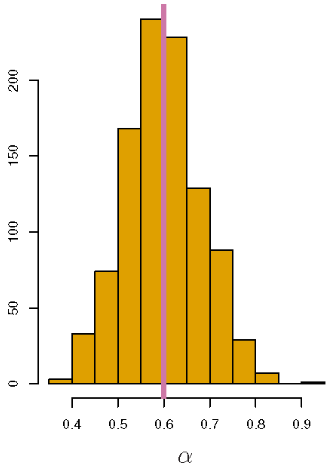


Bootstrap example with a sample of size 3

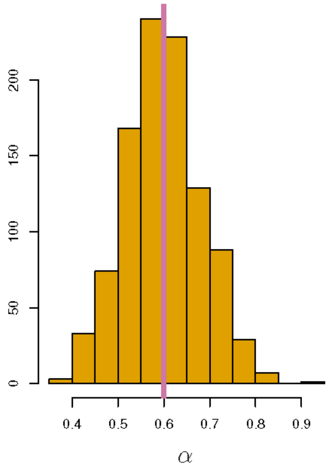
- Our original sample data Z includes 3 pairs of X and Y.
- We use the 1st bootstrap dataset Z^{*1} to produce a new bootstrap estimate $\widehat{\alpha}^{*1}$.
- This procedure is repeated B times (say, $B = 1000$) to produce B different bootstrap datasets $Z^{*1}, Z^{*2}, \ldots, Z^{*B}$ and B corresponding estimates $\widehat{\alpha}^{*1}, \widehat{\alpha}^{*2}, \ldots, \widehat{\alpha}^{*B}$.
- We then estimate the standard error of these bootstrap estimates via

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

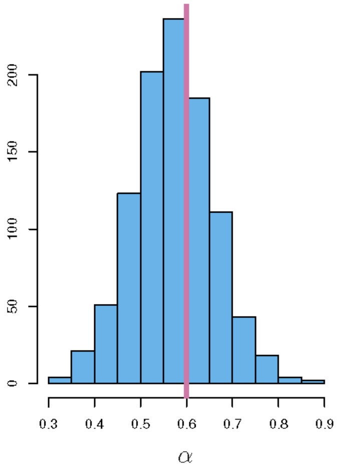




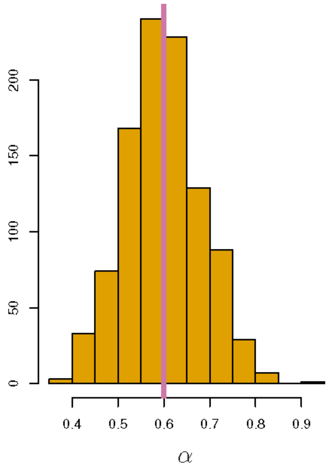
A histogram of the estimates of α obtained by generating 1000 simulated datasets from the true population.

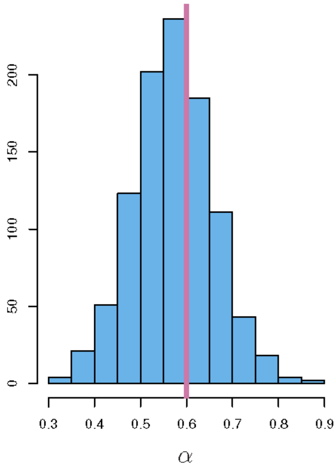


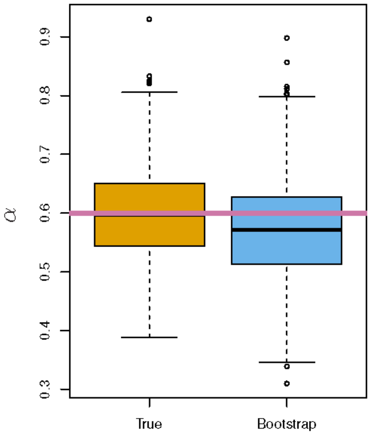
A histogram of the estimates of α obtained by generating 1000 simulated datasets from the true population.



A histogram of the estimates of α obtained from 1000 bootstrap samples from a single dataset.







A histogram of the estimates of α obtained by generating 1000 simulated datasets from the true population.

A histogram of the estimates of α obtained from 1000 bootstrap samples from a single dataset.

The same two distributions from the left and center panels shown as boxplots.

General use of bootstrap

- Primarily used to obtain standard errors of an estimate.
- Also provides approximate confidence intervals for a population parameter. For example, looking at the histogram in the middle panel of the previous slide, the 5% and 95% quantiles are 0.43 and 0.72, respectively. This represents an approximate 90% confidence interval for the true α.
- The above interval is called a *bootstrap percentile confidence interval*. It is the simplest method (among many approaches) for obtaining a confidence interval from the bootstrap.

General use of bootstrap

- Primarily used to obtain standard errors of an estimate.
- Also provides approximate confidence intervals for a population parameter. For example, looking at the histogram in the middle panel of the previous slide, the 5% and 95% quantiles are 0.43 and 0.72, respectively. This represents an approximate 90% confidence interval for the true α.
- The above interval is called a *bootstrap percentile confidence interval*. It is the simplest method (among many approaches) for obtaining a confidence interval from the bootstrap.
- In more complex data situations one needs to be careful about what the appropriate way to generate bootstrap samples is.
  - E.g. if data is a time series, we can't simply sample the observations with replacement.
  - We can instead create blocks of consecutive observations, and sample those with replacements. Then we paste together sampled blocks to obtain a bootstrap dataset.