Lecture 12: Resampling Methods (Textbook 5.2)

Nayel Bettache

Department of Statistical Science, Cornell University

Bootstrap

The bootstrap is mostly used to estimate the standard errors of some estimates.

We consider the following simple example.

- Suppose that 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.
- We wish to choose α to minimize the total risk, or variance, of our investment $\mathbb{V}(\alpha X + (1-\alpha)Y)$.
- 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 $\sigma_{XY} = Cov(X, Y)$.

Example

- We estimate σ_Y^2, σ_X^2 and σ_{XY} by the sample variance $\hat{\sigma}_Y^2$, $\hat{\sigma}_X^2$ and sample covariance $\hat{\sigma}_{XY}$ based on 100 data points $(x_1, y_1), ..., (x_{100}, y_{100})$.
- Then, we estimate α by

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

- How to estimate the variance of the estimator $\hat{\alpha}$?

A Non-practical Approach

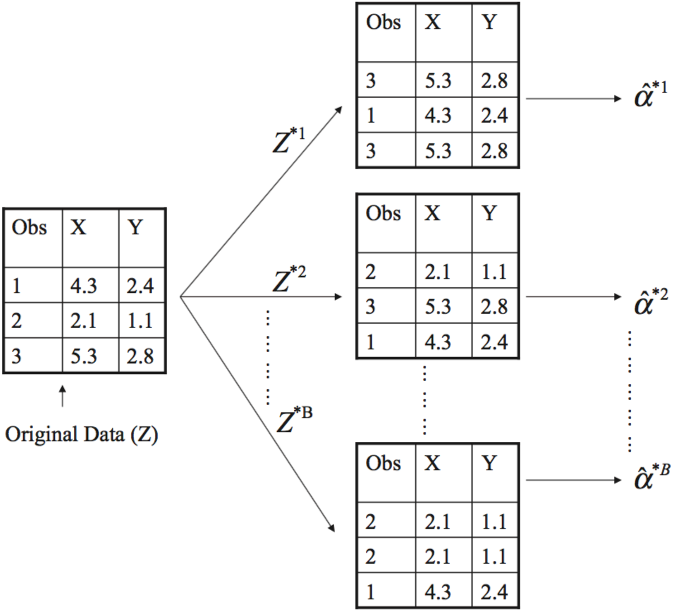
- If we know the distribution of X and Y (usually not true in reality), we can estimate the variance of the estimator $\hat{\alpha}$ by the following simulation based approach.
- We simulate 100 paired observations of X and Y and compute $\hat{\alpha}$. We repeat this procedure 1000 times, and get $\hat{\alpha}_1,\ldots,\hat{\alpha}_{1000}$.
- We estimate $\mathbb{V}(\hat{\alpha})$ by

$$\frac{1}{1000-1}\sum_{r=1}^{1000}(\hat{\alpha}_r - \bar{\alpha})^2, \text{ where } \bar{\alpha} = \frac{1}{1000}\sum_{r=1}^{1000}\hat{\alpha}_r.$$

Bootstrap

- The procedure outlined above cannot be applied, because for real data we cannot generate new samples from the original population.
- However, the bootstrap approach allows us to use a computer to mimic the process of obtaining new data sets, so that we can estimate the variability of our estimate without generating additional samples.
- Rather than repeatedly obtaining independent data sets from the population, we instead obtain distinct data sets by repeatedly sampling observations from the original data set **with replacement**.
- Each of these 'bootstrap data sets' 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


Obs X Y
1 4.3 2.4
2 2.1 1.1
3 5.3 2.8
Original Data (Z)
Z^{*1}
Z^{*2}
Z^{*B}
Obs X Y
3 5.3 2.8
1 4.3 2.4
3 5.3 2.8
Obs X Y
2 2.1 1.1
3 5.3 2.8
1 4.3 2.4
Obs X Y
2 2.1 1.1
2 2.1 1.1
1 4.3 2.4
$\hat{\alpha}^{*1}$
$\hat{\alpha}^{*2}$
$\hat{\alpha}^{*B}$

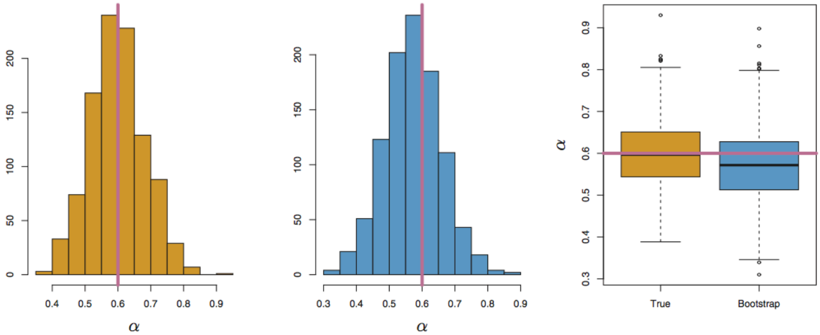
Bootstrap

How to use bootstrap to estimate $\mathbb{V}(\hat{\alpha})$?

- We denote the first bootstrap data set by Z^{*1}, and use Z^{*1} to form an estimate of α, denoted by $\hat{\alpha}^{*1}$.
- This procedure is repeated B times for some large value of B (say 1000), in order to produce B different bootstrap data sets, Z^{*1},...,Z^{*B} and B corresponding α estimates $\hat{\alpha}^{*1}$,...,$\hat{\alpha}^{*B}$.
- We estimate $\mathbb{V}(\hat{\alpha})$ by the sample variance of $\hat{\alpha}^{*1}$,...,$\hat{\alpha}^{*B}$:

$$\frac{1}{B-1}\sum_{r=1}^{B}(\hat{\alpha}^{*r}-\bar{\alpha}^{*})^2, \text{ where } \bar{\alpha}^{*}=\frac{1}{B}\sum_{r=1}^{B}\hat{\alpha}^{*r}.$$

Example



Left: A histogram of the estimates of α obtained by generating 1,000 simulated data sets from the true population.

Center: A histogram of the estimates of α obtained from 1,000 bootstrap samples from a single data set.

Right: The boxplots for estimates of α displayed in the left and center panels.