# STA 314: Statistical Methods for Machine Learning I

## Lecture - Bootstrap

Xin Bing

Department of Statistical Sciences
University of Toronto

# Bootstrap

- Bootstrap is a widely used **resampling** (of the available data) approach!
- It can be used to assess the uncertainty of basically **any** statistical procedure.
  - For instance, it can be used to estimate the standard errors of the estimated coefficients of a linear model.
  - Much more generally, it can even estimate the whole distribution of the estimated coefficients of a linear model.
- Its validity is backed up by a very general theory!

# A simple example

- Suppose that we wish to invest 10K dollars in two financial assets that yield returns of $X$ and $Y$, respectively, where $X$ and $Y$ are random quantities.
- For any $\alpha \in [0, 1]$, we will invest a fraction $\alpha$ of our money in $X$, and will invest the remaining $(1 - \alpha)$ in $Y$.

$$\Big[\alpha X + (1 - \alpha)Y\Big] \times 10{,}000$$

- We wish to choose $\alpha$ to minimize the total risk, or variance, of our investment, that is,

$$\min_{\alpha \in [0,1]} \mathrm{Var}(\alpha X + (1 - \alpha)Y).$$

# Example

- One can show that the value of $\alpha$ that minimizes the risk is given by

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

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

- If we have past observations $(x_1, y_1), ..., (x_{100}, y_{100})$, we can estimate $\alpha$ 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 $\hat{\alpha}$?

# An Oracle Approach

If we know the distribution of $X$ and $Y$ (usually unknown in reality), we can estimate the variance of the estimator $\hat{\alpha}$ by the following strategy.

- We simulate 100 paired observations of $X$ and $Y$ and compute

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

- We repeat this procedure 1000 times, and get $\hat{\alpha}_1, \ldots, \hat{\alpha}_{1000}$.

- We estimate $\text{Var}(\hat{\alpha})$ by

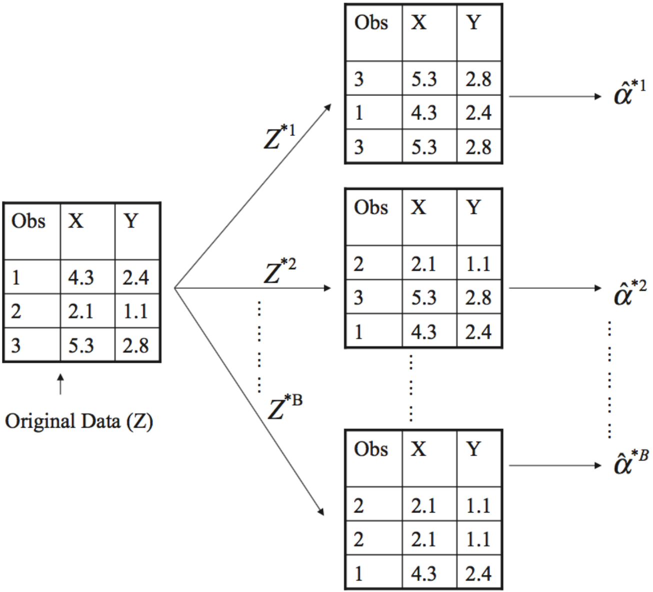
$$\frac{1}{1000-1}\sum_{r=1}^{1000}(\hat{\alpha}_r - \bar{\alpha})^2, \quad \text{where} \quad \bar{\alpha} = \frac{1}{1000}\sum_{r=1}^{1000}\hat{\alpha}_r.$$

**Q:** Is this feasible in practice?

# Bootstrap

- The bootstrap approach re-samples from the **original data set** to mimic the process of obtaining new data sets in order to quantifty the uncertainty of a given procedure.
- Specifically, for a specified $B$ (for instance, $B = 1000$) number of repetitions, we repeatedly sample **the same amount of observations** from the original data set **with replacement**.
- As a result, data set from bootstrap might contain some observations more than once, or zero time.

# Simple illustration of Bootstrap

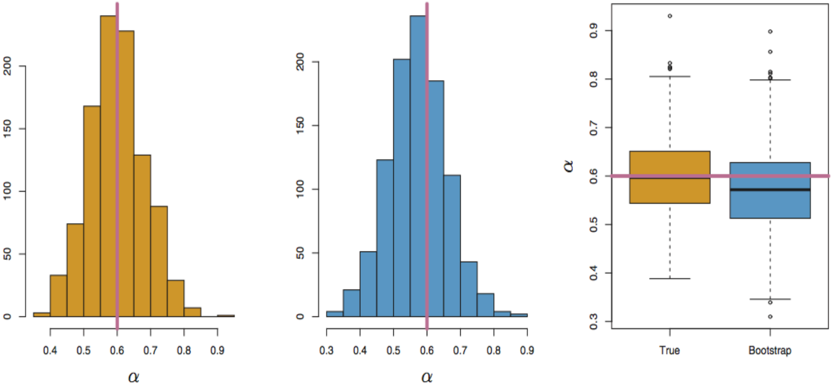


Now let us apply bootstrap to estimate $\text{Var}(\hat{\alpha})$:

- We denote the first bootstrap data set by $Z^{*1}$, and use $Z^{*1}$ to construct the estimate of $\alpha$, denoted by $\hat{\alpha}^{*1}$.
- This procedure is repeated $B$ (say, $B = 1000$): specifically we simulate $B$ different bootstrap data sets, $Z^{*1}, \ldots, Z^{*B}$ and $B$ corresponding estimates $\hat{\alpha}^{*1}, \ldots, \hat{\alpha}^{*B}$.
- We estimate $\text{Var}(\hat{\alpha})$ by the sample variance of $\hat{\alpha}^{*1}, \ldots, \hat{\alpha}^{*B}$:

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

# Example




- Left: The histogram of estimates of $\alpha$ obtained by generating 1,000 simulated data sets from the true population.
- Center: The histogram of estimates of $\alpha$ obtained from 1,000 bootstrap samples from a single data set.
- Right: Boxplots for estimates of $\alpha$ displayed in the left and center panels.

# Bootstrap for quantifying the uncertainty of the OLS estimator

Given the data $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, the OLS gives

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

Statistical property of $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ consists of

- its mean
- its covariance
  - *Recall that analyses of $\hat{\boldsymbol{\beta}}$ such as its mean and covariance are only available under linear model assumption.*
- its higher moments
- its whole distribution

Bootstrap can be used to estimate **all above**!

# Bootstrap for quantifying the uncertainty of the OLS estimator

Given the data $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, the OLS gives

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

For $b = 1, \ldots, B$,

1. obtain the bootstrap sample $D^b = (\mathbf{X}^b, \mathbf{y}^b)$
2. compute $\hat{\boldsymbol{\beta}}^b = (\mathbf{X}^{b\top} \mathbf{X}^b)^{-1} \mathbf{X}^{b\top} \mathbf{y}^b$.

Now for $\hat{\beta}_j$, the bootstrap estimates $\hat{\beta}_j^1, \ldots, \hat{\beta}_j^B$ serve as "samples" of $\hat{\beta}_j$.

# Bootstrap for quantifying the uncertainty of the OLS estimator

For instance,

- the mean of $\hat{\beta}_j$ can be estimated by

$$\frac{1}{B}\sum_{b=1}^{B}\hat{\boldsymbol{\beta}}_j^b.$$

- the variance of $\hat{\beta}_j$ can be estimated by

$$\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\boldsymbol{\beta}}_j^b - \frac{1}{B}\sum_{b=1}^{B}\hat{\boldsymbol{\beta}}_j^b\right)^2.$$

- You can also estimate quantiles of the distribution of $\hat{\beta}_j$.
- In fact, you can estimate the whole distribution of $\hat{\beta}_j$.