Module 5: Resampling Methods

Lecture 11
Feb 8th, 2023
Ch 5.1.3-4:

Recap

MICHIGAN STATE UNIVERSITY

- LOOCV $\quad \mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{MSE}_i$
- Leverage – outlier
- K-fold CV $\quad \mathrm{CV}_{(K)} = \sum_{k=1}^{K} \frac{n_k}{n} \mathrm{MSE}_k$

$K = n$

K-fold Cross-validation

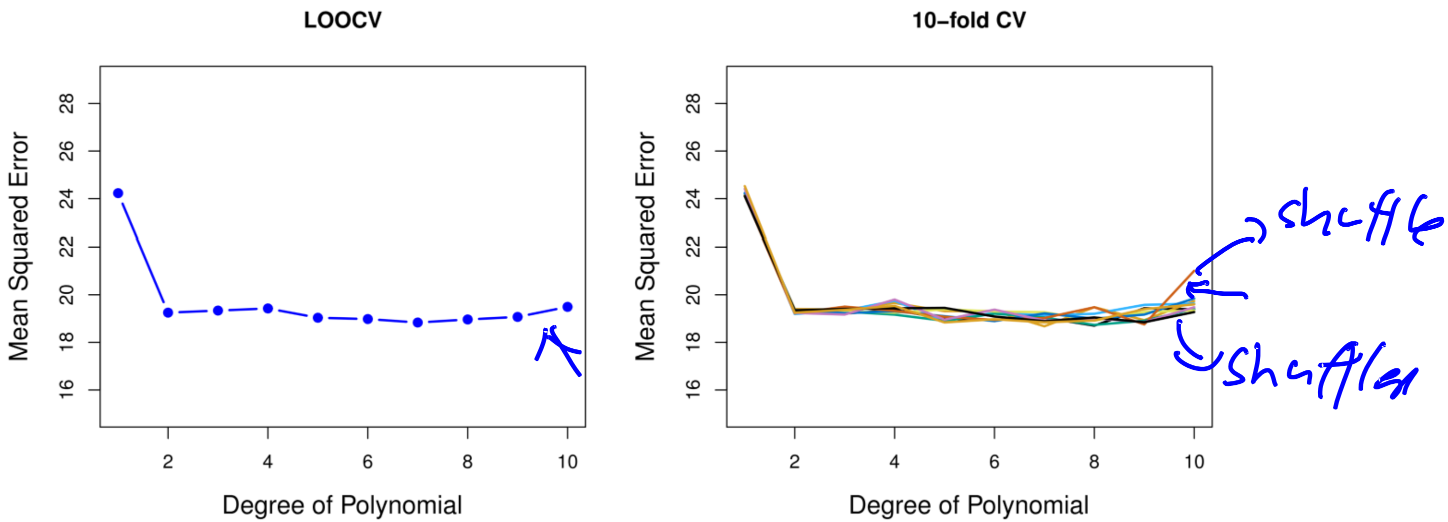
MICHIGAN STATE UNIVERSITY

- Hybrid between validation set and LOOCV

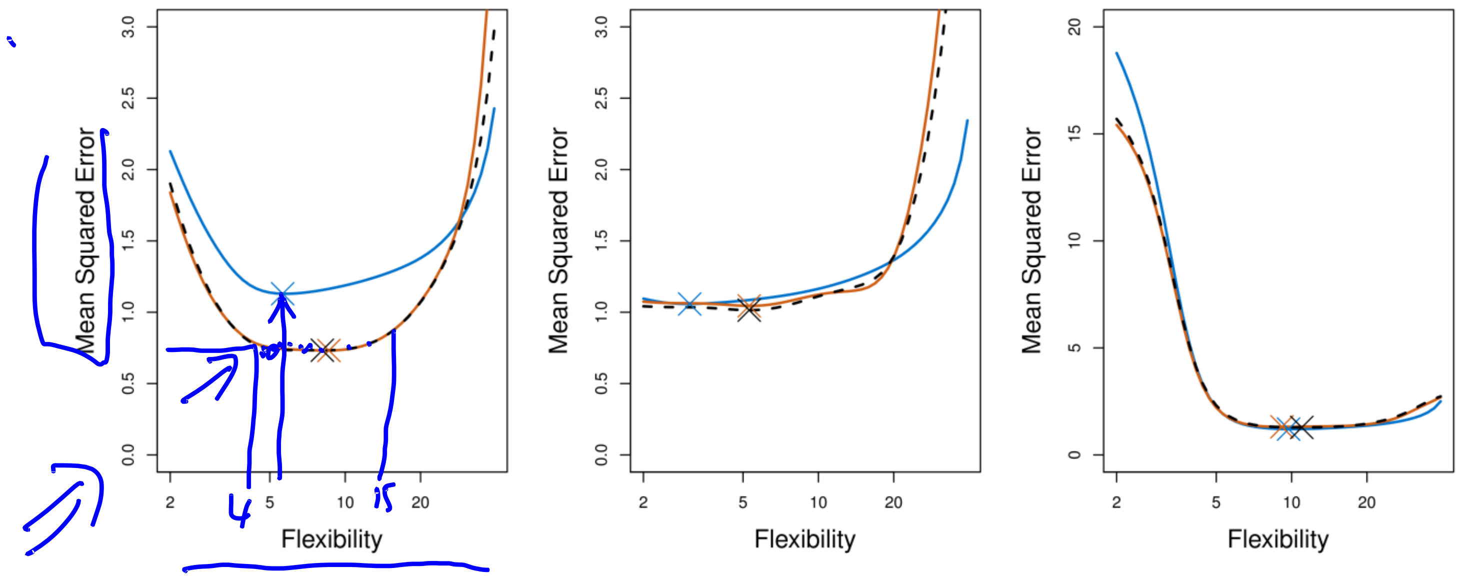
$$\mathrm{CV}_{(K)} = \sum_{k=1}^{K} \frac{n_k}{n} \mathrm{MSE}_k$$

Coding – Building k-fold CV

Cross-validation vs LOOCV



Simulated Results



Blue True error

Dashed LOOCV estimate

Orange 10-fold CV

CV for Classification

- We divide the data into *K* roughly equal-sized parts

Compute

$$\text{CV}_K = \sum_{k=1}^{K} \frac{n_k}{n} \text{Err}_k$$

where $\text{Err}_k = \sum_{i \in C_k} I(y_i \neq \hat{y}_i)/n_k$.

Example: GWAS

5000 >> 50

A T C G

Grace

Clark

Terene

Tao

Alan Turing

- Sample size n = 50, number of predictors (SNPs) p = 5000; Predictor heart attach after age of 60.
- Step 1: Select the top 100 predictors having the largest correlation with the class labels
- Step2: We then apply a classifier (logistic regression) using only these 100 predictors

How do we estimate the test set performance of this classifier?

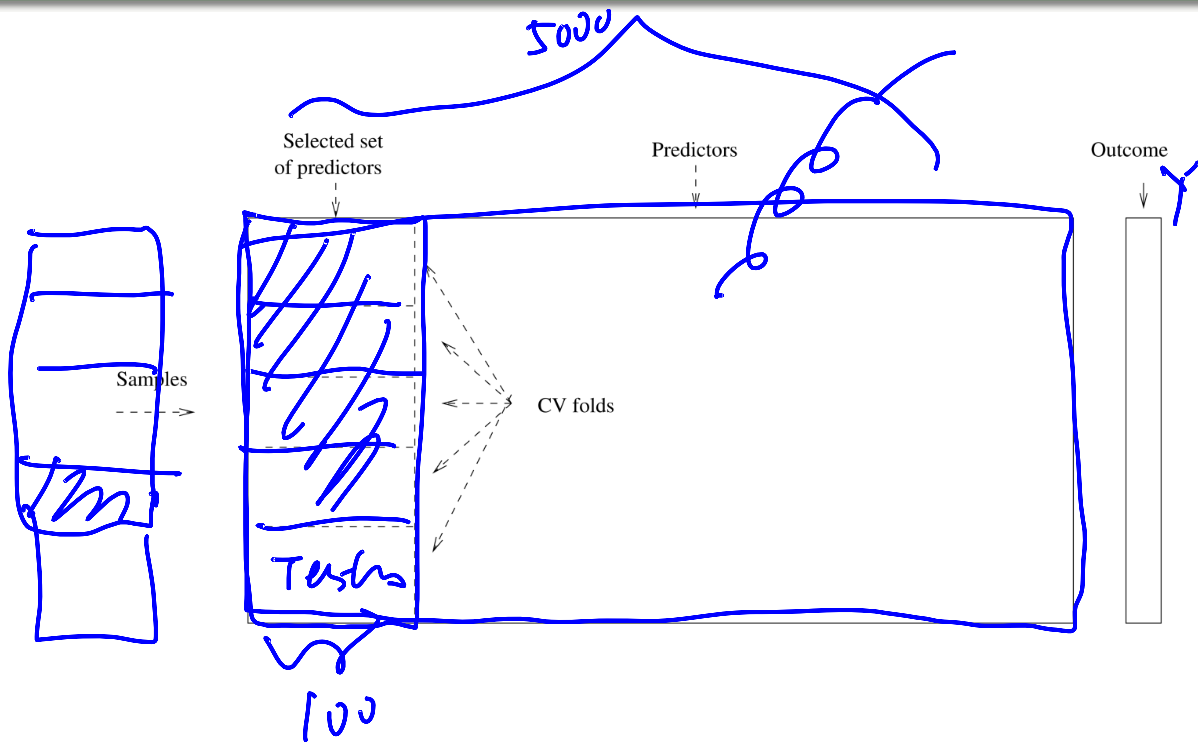
P1 ATCGATTA

P2 ATCGTTTA

SNP

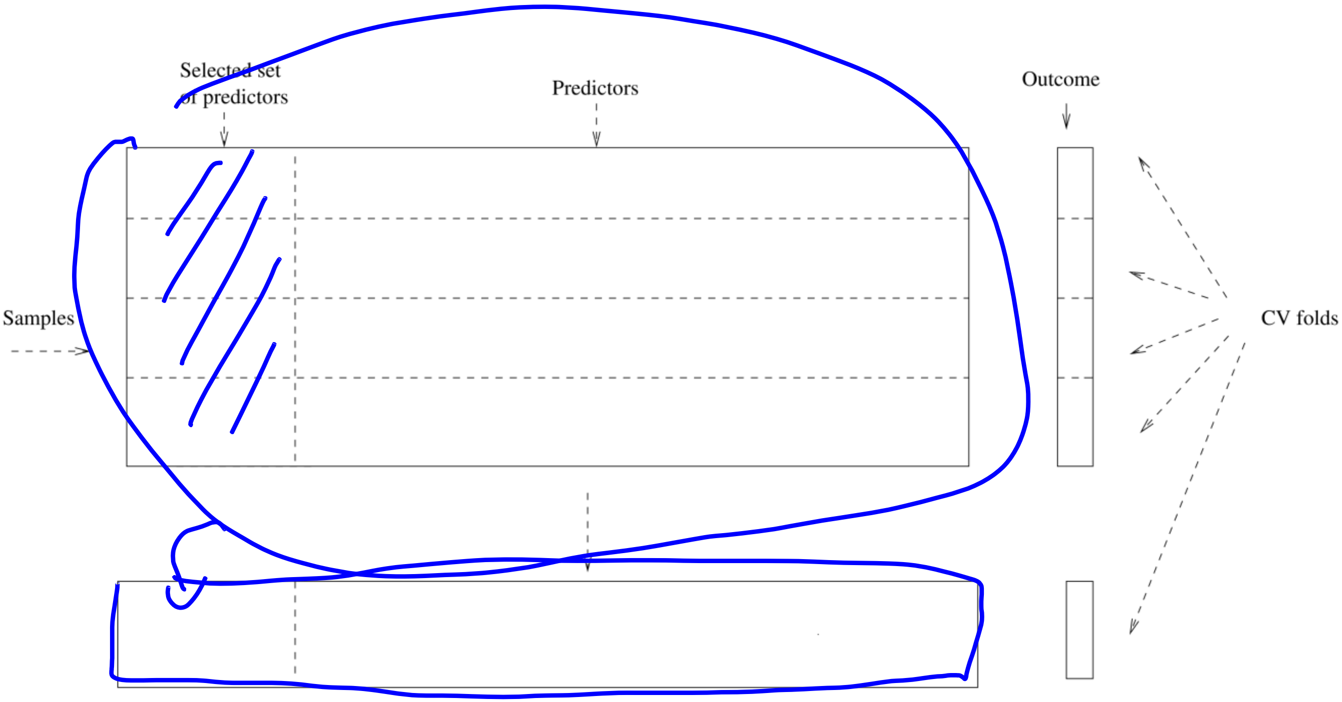
Example: GWAS

MICHIGAN STATE UNIVERSITY



No ! ! !

Example: GWAS ------Right Way



information leaky

Coding

Bootstrap

CV

- Bootstrap is used to quantify the uncertainty associated with an estimator or machine learning method.

  It can provide an estimate of the standard error of a coefficient. Then we can do hypothesis test and confidence interval.

$\boxed{\hat{\beta}}$? $\boxed{\hat{f}}$?

Bootstrap: Finance Example

MICHIGAN STATE UNIVERSITY

TSLA

KO

- To invest two stocks that yield return of X and Y, where X and Y are random.
- We will invest a fraction α of our money in X and the rest in Y.
- Assume both stock have the same average return over the years. What criteria should we use for allocate the investment?

$E(X) = E(Y)$

α

$Var(X) \gg Var(Y)$

Bootstrap： Finance Example

- We want to minimize the total risk or variance of our investment.

$$\boxed{\mathrm{Var}(\alpha X + (1-\alpha)Y)}.$$

$\mathrm{Var}(X)$

$\mathrm{Cov}(X, X)$

- The solution is

$X = Y$ $\quad$ $\mathrm{Var}(X+Y) \neq \mathrm{Cov}(X,Y)$

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

$\mathrm{Var}(2X) = 2^2\,\mathrm{Var}(X)$

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

$(X+Y) = (1 \ 1)\begin{pmatrix} X \\ Y \end{pmatrix}$ $\quad$ $\mathrm{Var}\left((1 \ 1)\begin{pmatrix} X \\ Y \end{pmatrix}\right)$

$= (1 \ 1)\,\mathrm{Var}\left(\begin{pmatrix} X \\ Y \end{pmatrix}\right)\begin{pmatrix} 1 \\ 1 \end{pmatrix}$

$= (1 \ 1)\begin{bmatrix} \mathrm{Var}(X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(X,Y) & \mathrm{Var}(Y) \end{bmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$

= Var(X) + Var(Y) + 2 Cov(X, Y)

Finance Example: Simulations

MICHIGAN STATE UNIVERSITY

- But the values of σ_X^2, σ_Y^2, and σ_{XY} are unknown.
- We can compute estimates for these quantities, $\hat{\sigma}_X^2$, $\hat{\sigma}_Y^2$, and $\hat{\sigma}_{XY}$, using a data set that contains measurements for X and Y.
- We can then estimate the value of α that minimizes the variance of our investment using

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

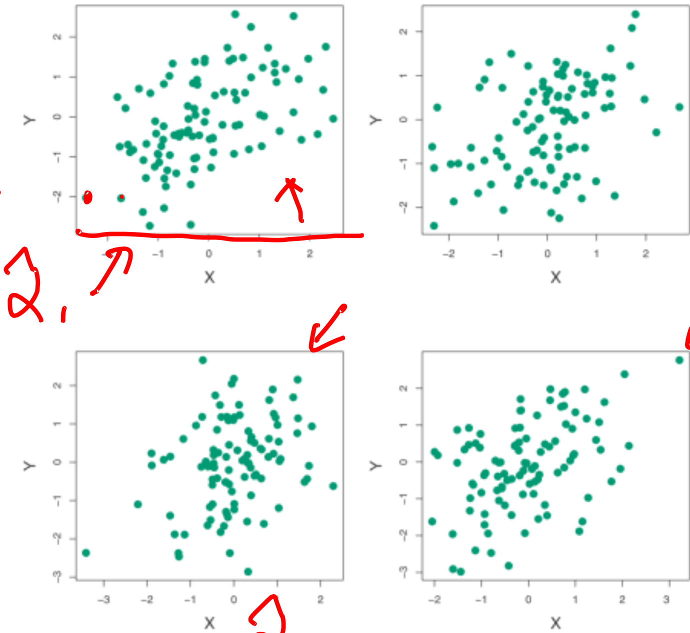
Finance Example: Simulations

MICHIGAN STATE UNIVERSITY

100

$\hat{\sigma}_X^2$, $\hat{\sigma}_Y^2$, $\hat{\sigma}_{XY}^2 \Rightarrow \hat{\alpha}_1$

$\hat{\alpha}_2$



$\hat{\alpha}_3$

FIGURE 5.9. *Each panel displays* 100 *simulated returns for investments* X *and* Y. *From left to right and top to bottom, the resulting estimates for* α *are* 0.576, 0.532, 0.657, *and* 0.651.

Finance Example: Simulations

MICHIGAN STATE UNIVERSITY

- To estimate the standard deviation of $\hat{\alpha}$, we repeated the process of simulating 100 paired observations of X and Y, and estimating α 1,000 times.
- We thereby obtained 1,000 estimates for α, which we can call $\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_{1000}$.

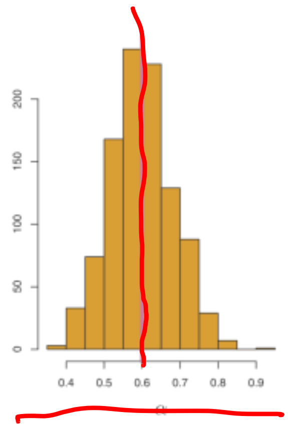
- For these simulations the parameters were set to $\sigma_X^2 = 1, \sigma_Y^2 = 1.25$, and $\sigma_{XY} = 0.5$, and so we know that the true value of α is 0.6 (indicated by the red line).

- The mean over all 1,000 estimates for α is

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

very close to $\alpha = 0.6$, and the standard deviation of the estimates is

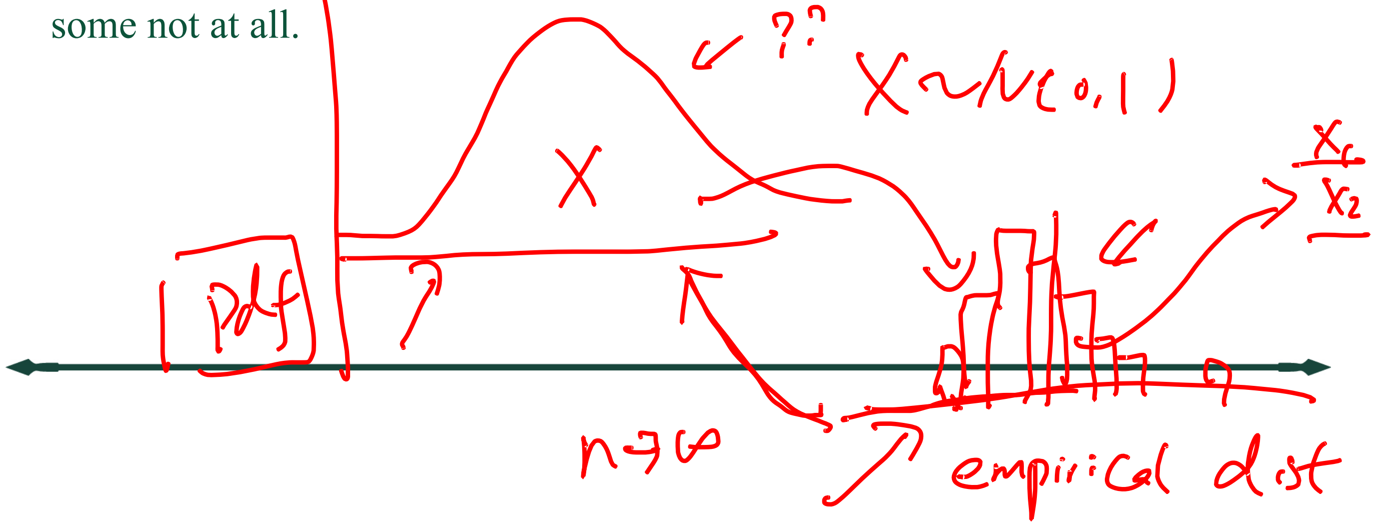
$$\sqrt{\frac{1}{1000 - 1} \sum_{r=1}^{1000} (\hat{\alpha}_r - \bar{\alpha})^2} = 0.083.$$



- This gives us a very good idea of the accuracy of $\hat{\alpha}$: $\mathrm{SE}(\hat{\alpha}) \approx 0.083$.

Finance Example: Reality

- The procedure outlined above cannot be applied, because for real data we cannot generate new samples from the original population.
- The bootstrap approach allows us to use a computer to mimic the process of obtaining new data sets.
- Rather than repeatedly sampling from the population, we instead obtain distinct data sets by repeatedly sampling observations from the original data set with replacement. n
- Each of these "bootstrap data sets" is created by sampling with replacement, and is the same size as our original dataset. Some observations may appear more than once in a given bootstrap data set and some not at all.



Boostrap Illustration

n:3

Populat...

Original Data (Z)

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

n=3

Z^{*1}

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

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

Z^{*2}

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

$\rightarrow \hat{\alpha}^{*2}$

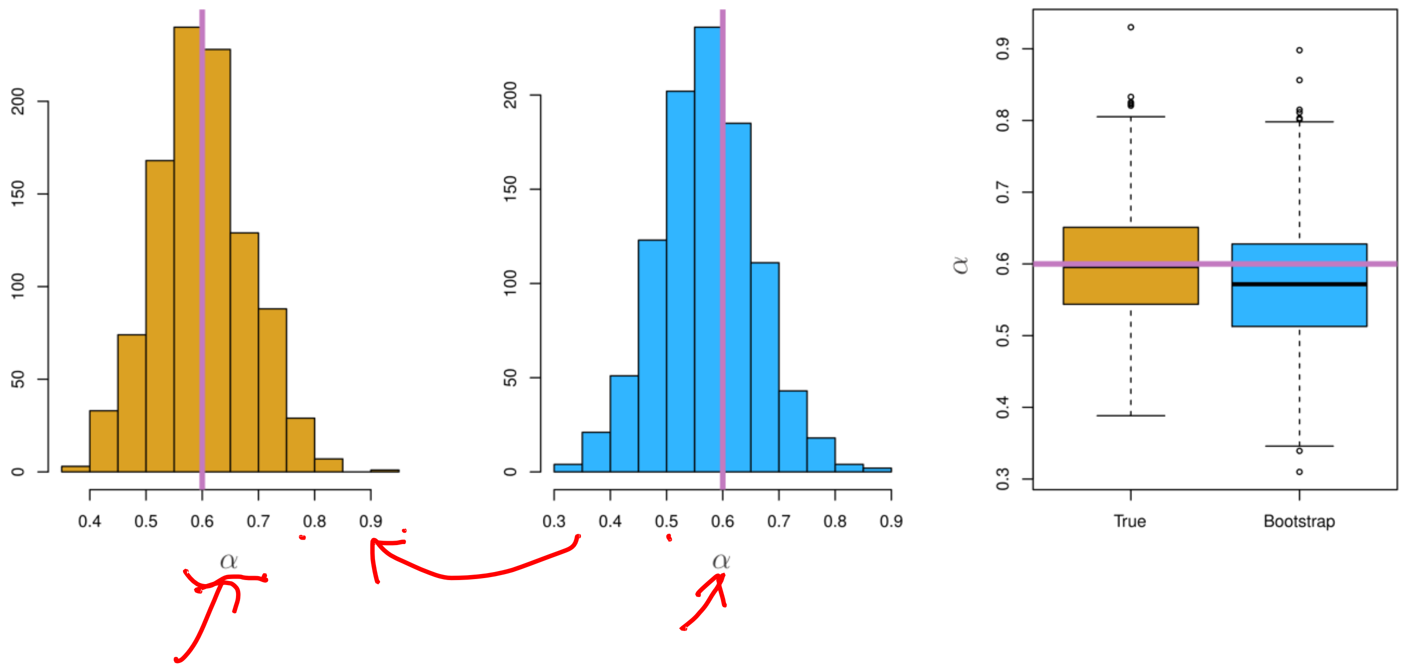
⋮

Z^{*B}

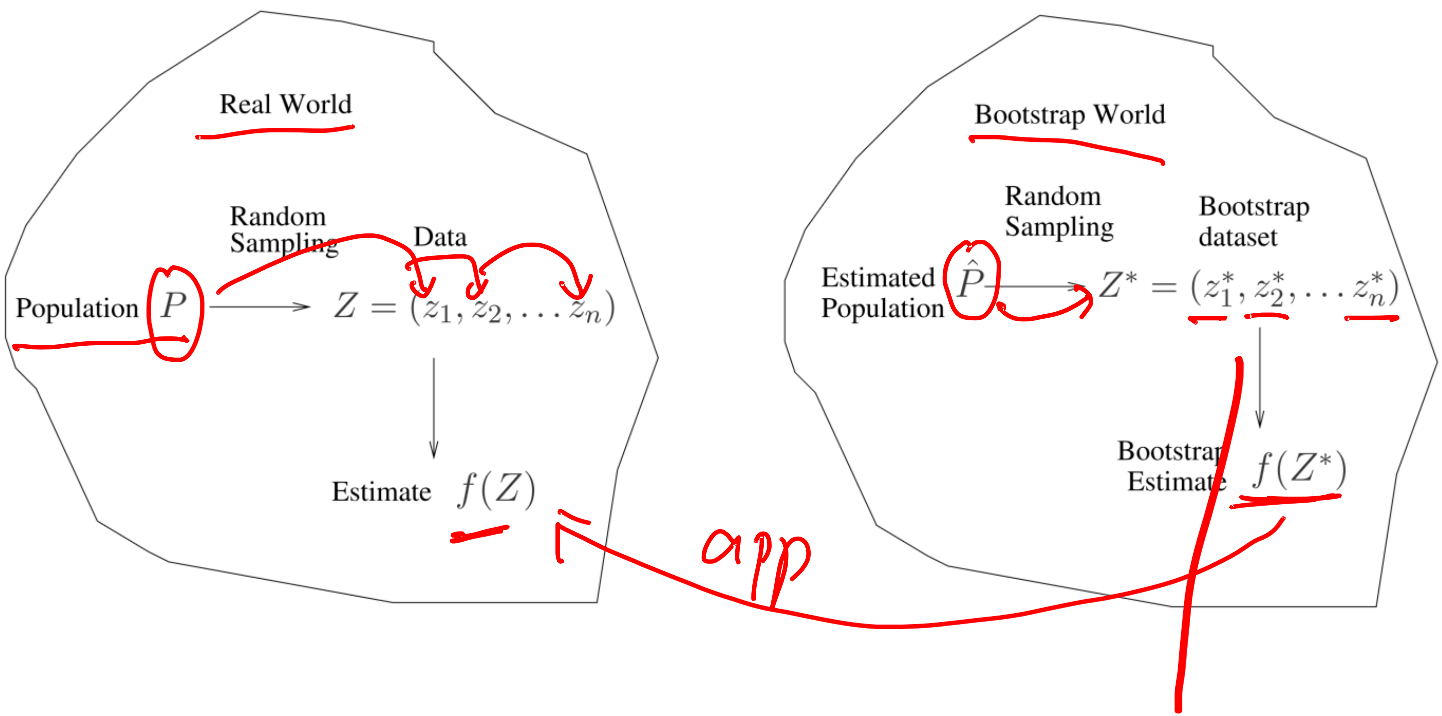
Obs	X	Y
2 ✓	2.1	1.1
2 ✓	2.1	1.1
1 ✓	4.3	2.4

$\rightarrow \hat{\alpha}^{*B}$

Boostrap Results



A General Picture for the Bootstrap



Boostrap for Prediction Error?

MICHIGAN STATE UNIVERSITY

Bonus Quiz

- Can we use one boostrap dataset as training and the original data set as test set? ?

B_1

n 100

Boostrap

n: 100

training

?

testing

Bonus Quiz 13

MICHIGAN STATE UNIVERSITY

(20 pts) For a classification problem with $K = 2$ ($Y \in \{0, 1, 6\}$), we know the oracle classifier is

$$C(x) = j, \quad \text{if } p_j(x) = \max\{p_0(x), p_1(x)\},$$

which is based on the loss with equal weight for Type I and II error. If we know Type I erorr will cost \$1000 while Type II error will cost \$3000. Derive the new oracle classifier which minimizes this cost.

classification logistic regression

Bonus Quiz 15

MICHIGAN STATE UNIVERSITY

- When sample size n is large, we know a bootstrap dataset will contain $1 – e^{-1} = 63.2\%$ of original data. Write a code to demonstrate it using $n = 1000000$

① Supervised
- regression → ① KNN ② Linear regression
- Classification ① KNN ② Logistic reg
  - → ~~Test~~ the best model?

Loss function

② unsupervised

Bias-Variance trade-off

$\hat{\beta} = ?$ SE($\hat{\beta}$)? CI, HT

Linear regression ←

X is qualitative, interaction terms, beyond linearity?