IFI 9000 Analytics Methods
Resampling and Boosting

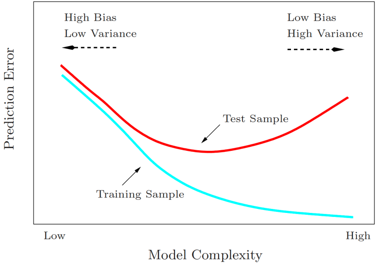
by **Houping Xiao**

Spring 2021

Georgia State University | J. MACK ROBINSON COLLEGE OF BUSINESS

Review of Cross Validation

- As mentioned earlier, model selection based on the RSS or R^2 statistics can be misleading, since the training error is not a good representative of the actual test error



- Instead through a process of splitting the data into training and validation sets, we were able to use LOOCV or K-fold CV as estimates of the test error
- The K-fold cv is more desirable estimate, computationally and statistically

Bootstrap

Bootstrap

- The bootstrap is a flexible and very powerful statistical tool that can be used to **quantify the uncertainty** with a given estimator or statistical learning method
- It can provide an estimate of the standard error of a coefficient, or a **confidence interval for that coefficients**, regardless of how complex the derivation of that coefficient is

Bootstrap via an Example

- Lets explain bootstrap via an example, **Best investment allocation**
- Suppose that we wish to invest a fixed sum of money in two financial assets that yield returns of X and Y (random quantities)
- We will invest α shares in X, and will invest the remaining $1 - \alpha$ in Y
- To minimize the risk, we want to minimize $var(\alpha X + (1 - \alpha)Y)$
- We can show that the minimizer is

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

where $\sigma_{XY}^2 = cov(X, Y)$ and $\sigma_Y^2 = var(Y)$

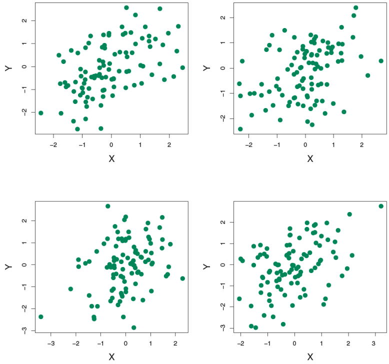
Bootstrap via an Example

- In practice, σ_X^2, σ_Y^2 and σ_{XY} are unknown
- Suppose we are given a data set containing pairs of X and Y. We can estimate σ_X^2, σ_Y^2 and σ_{XY} from the sample set and get an estimate $\hat{\alpha}$ for the optimal share
- Ideally, we can generate these sample sets many times, and estimate an $\hat{\alpha}$ for each and look into the histogram
- However, in practice we only have one sample set to use
- Bootstrap yet allows us to generate good estimates of α **using only one sample set**

Bootstrap via an Example

To see how nicely bootstrap works, let's compare its outcome with the case that α is generated from many synthetic sample generations

- We generate 1,000 sample sets each containing 100 pairs of X and Y
- For the synthetic data generated $\sigma_X^2 = 1, \sigma_Y^2 = 1.25$ and $\sigma_{XY}^2 = 0.5$ which yield an optimal value of $\alpha = 0.6$



Bootstrap via an Example

- To get the left panel we generate 1,000 synthetic sample sets, for each obtain $\hat{\alpha}$ and plot the histogram and calculate

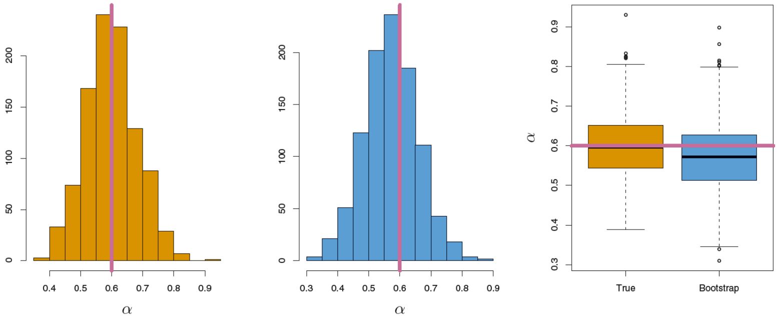
$$\bar{\alpha} = \frac{1}{1000} \sum_{i=1}^{1} 000\hat{\alpha}_i = 0.5996$$

$$SE(\alpha) = \sqrt{\frac{1}{1000-1} \sum_{i=1}^{1} 000(\hat{\alpha}_i - \bar{\alpha})^2} = 0.083$$

- For the bootstrap we only use one of the sample sets and regenerate new sample set by **sampling with replacement**
- Surprisingly, the results are very close to $\alpha = 0.6$

Bootstrap via an Example

- Surprisingly, the results are very close to $\alpha = 0.6$

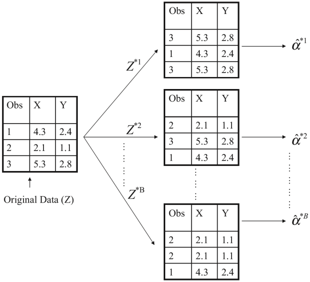


Left: A histogram of the estimates of α obtained by generating 1,000 simulated data sets from the true population. **Center**: A histogram of the estimates of obtained from 1,000 bootstrap samples from a single data set. **Right**: The estimates of displayed in the left and center panels are shown as boxplots. In each panel, the pink line indicates the true value of α.

Bootstrap General Framework

- Suppose a black-box calculates $\hat{\alpha}$ from a sample set, e.g., a coefficient in linear or logistic regression
- We are interested in estimating the variability of $\hat{\alpha}$ without examining many new sample sets
- Denoting the first bootstrap data set by Z^{*1}, we use Z^{*1} to produce a new bootstrap estimate for α, which we call $\hat{\alpha}^{*1}$
- This procedure is repeated B (say 100 or 1,000) times, in order to produce B different bootstrap data sets, $Z^{*1}, Z^{*2}, \cdots, Z^{*B}$, and the corresponding α estimates $\hat{\alpha}^{*1}, \cdots, \hat{\alpha}^{*B}$

Bootstrap General Framework



- We estimate the standard error (SE) of these bootstrap estimates using the formula

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

- This serves as an estimate of the standard error of α estimated from the original data set

The End