COM 525000 – Statistical Learning

Lecture 5 – Resampling Methods

Y.-W. Peter Hong

Resampling Methods

- **Resampling** refers to the repeated drawing of samples from a training set and refitting a model of interest on each set of samples.
  - ➔ Used to obtain additional information about the fitted model.
  - ➔ E.g., for estimating the test error (for model assessment and selection), or to estimate the variability of coefficient estimates.
- Two common approaches:
  - – **Cross-validation**
  - – **Bootstrap**

Training versus Test Error (1/2)

- The **training error** is the average error between the fitted and true responses of data points in the training set. That is,

$$\overline{\mathrm{err}} = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{f}(x_i;\mathcal{D})) \quad \left(\overset{\text{e.g.}}{=} \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i;\mathcal{D}))^2 \right)$$

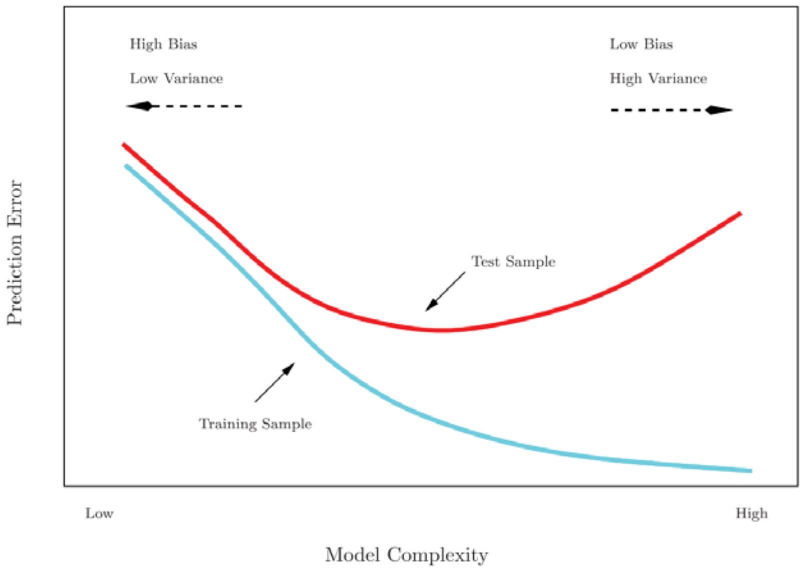
  where $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ is the available data.

- The **test error** (or, generalization error) is the average error of predicting the response on a *new observation*, i.e.,

$$\mathrm{Err} = E[L(Y, \hat{f}(X;\mathcal{D}))] \quad \left(\overset{\text{e.g.}}{=} E[(Y - \hat{f}(X;\mathcal{D}))^2] \right)$$

  where the expectation is taken over X, Y and $\mathcal{D}$.

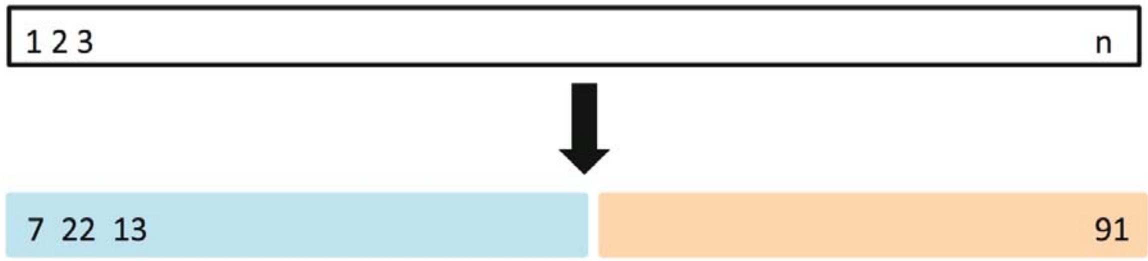
Training versus Test Error (2/2)



- Test set is often unavailable and, thus, the actual test error is often unknown. (➔It must be estimated!)
- **Key Idea:** Hold out a subset of the available data for testing later on, and train on the remaining subset.

The Validation Set Approach

- **The Validation Set Approach:**
  - Split into a training set $\mathcal{D}_{\text{train}}$ and a validation set $\mathcal{D}_{\text{val}}$.

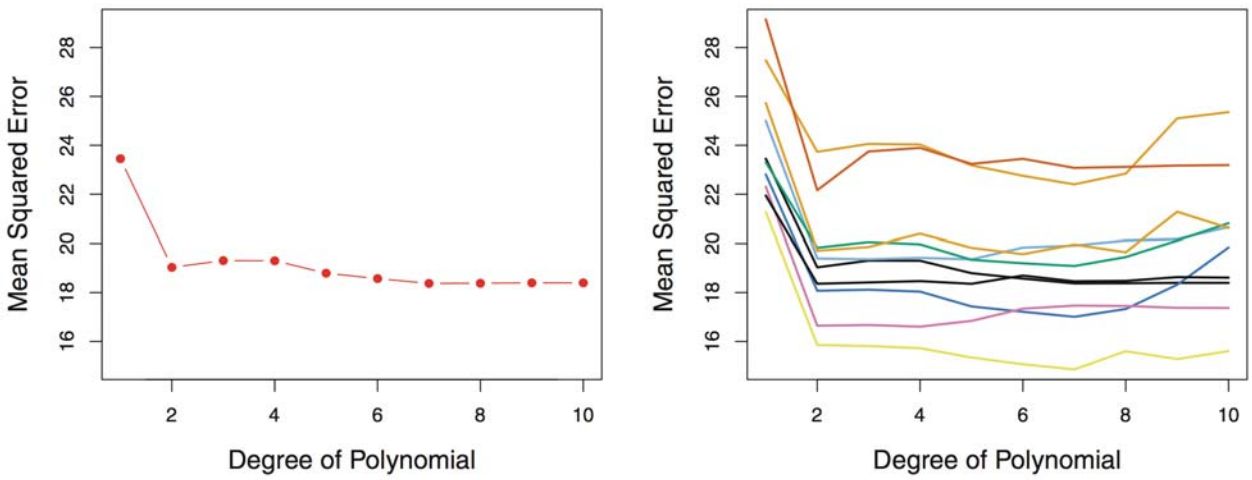


  - Train (or fit) the model on the training set, and predict the responses for observations in the validation set.

➔The validation set error rate provides an estimate of the test error rate.

Example: Auto Data Set

- X : horsepower, Y : mpg
- 392 observations=196 training set+196 validation set

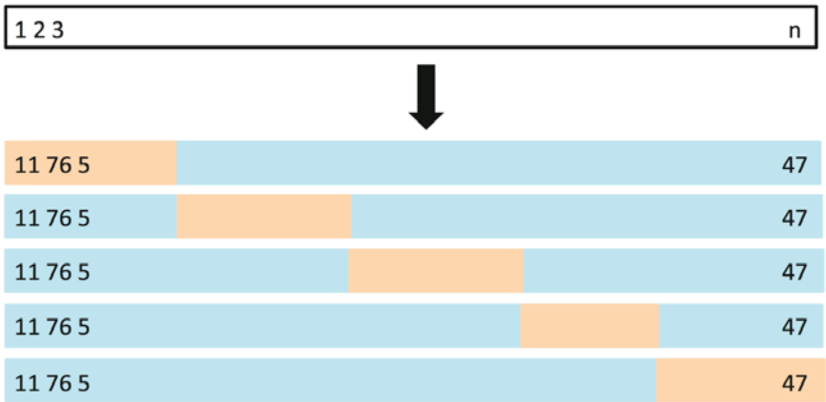


➔ The estimate of the test error is highly variable.

➔ The validation set error rate tend to *overestimate* the test error rate for the model fit on the entire data.

k-Fold Cross-Validation

- **k-Fold Cross-Validation:**
  – Split the data into k groups (or folds) $\mathcal{D}_1, \ldots, \mathcal{D}_k$ of size $n_1, \ldots, n_k$ (e.g., $n_1 = \cdots = n_k = \frac{n}{k}$).



  – Use the jth fold $\mathcal{D}_j$ as the validation set and the remaining $k-1$ folds $\mathcal{D} \setminus \mathcal{D}_j$ as the training set.
  – Repeat for $j=1, \ldots, k$ to get MSEs $\mathrm{MSE}_1, \ldots, \mathrm{MSE}_k$.

MSE of k-Fold CV

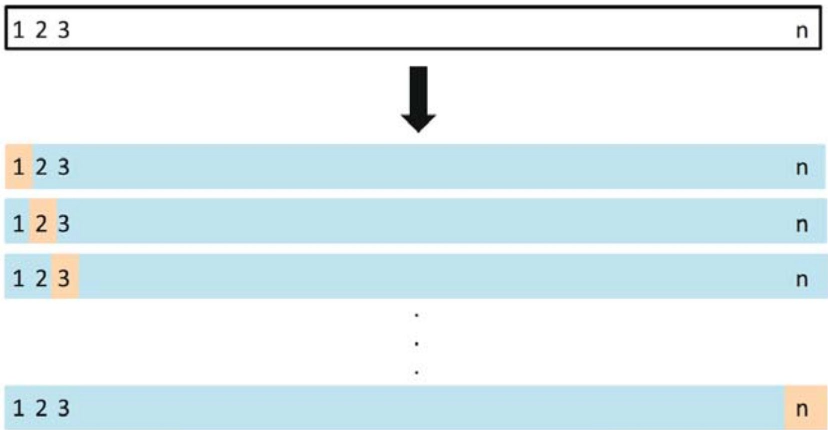
- The overall MSE is

$$\mathrm{CV}_{(k)} = \sum_{j=1}^{k} \frac{n_j}{n} \mathrm{MSE}_j \left(\overset{\text{e.g.}}{=} \frac{1}{k} \sum_{j=1}^{k} \mathrm{MSE}_j, \ \text{for } n_j = \frac{n}{k}, \forall j \right)$$

➔ Typical values of k are 5 and 10.

Leave-One-Out Cross-Validation (LOOCV)

- **Leave-one-out cross-validation (LOOCV)** is a special case of k-fold CV where $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ is split into n groups $\mathcal{D}_1 = \{(x_1, y_1)\}, \ldots, \mathcal{D}_n = \{(x_n, y_n)\}$.



➔Less bias (since larger training sets are used).

➔Computationally expensive.

Special Case of LOOCV MSE

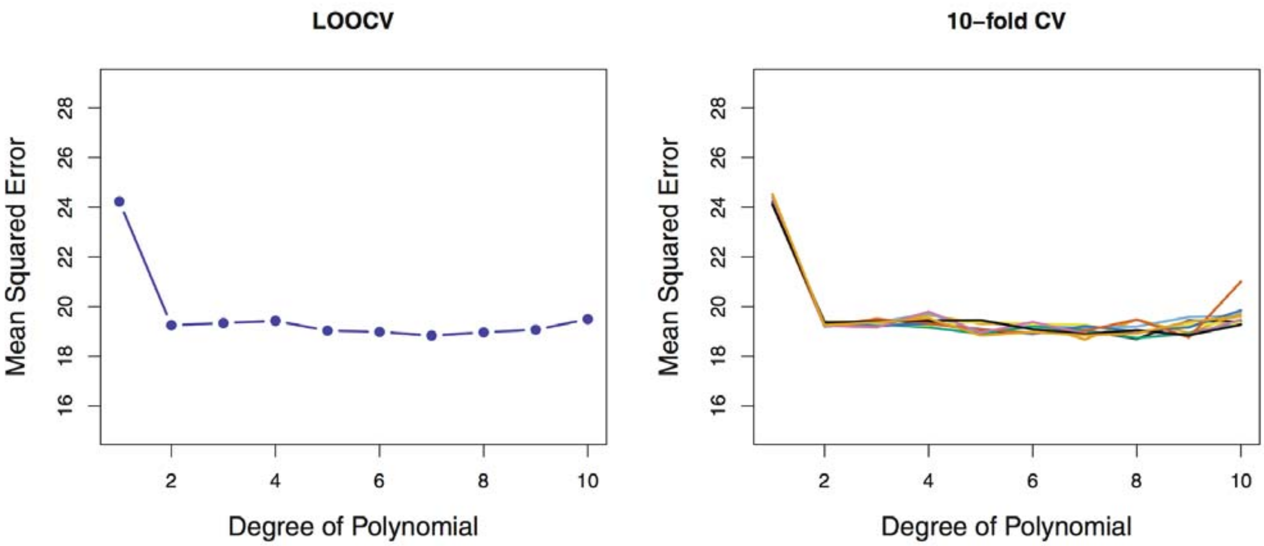
- **Special Case:** For least squares linear (or polynomial) regression, the cost of LOOCV can be computed as

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2$$

where $h_i \triangleq \{\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\}_{ii} = x_i^T(\mathbf{X}^T\mathbf{X})^{-1}x_i$.

LOOCV vs k-Fold CV

- **Revisit Auto Data Set Example:**

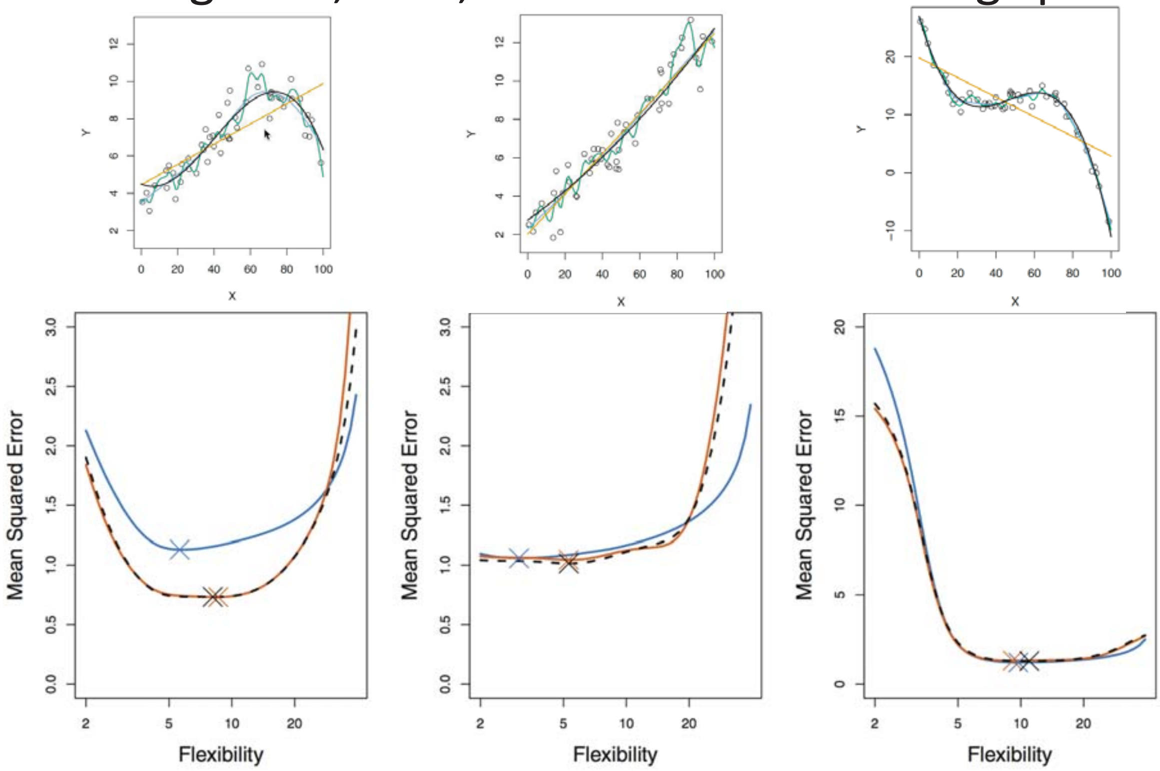


➔LOOCV yields a deterministic test error estimate since there is only one way to split the data set.

➔10-fold CV exhibits some variability for random splits.

LOOCV vs k-Fold CV (Simulated Example)

- Recall Figs. 2.9, 2.10, and 2.11 on smoothing splines.



Bias-Variance Tradeoff for k-Fold CV (1/2)

- CV methods are used to estimate the test error

$$E[(Y - \hat{f}(X;\mathcal{D}))^2]$$

where $\mathcal{D}$ is the available data set.

- In k-fold CV (with $n_j = n/k, \forall j$), this is estimated by

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{j=1}^{k}\frac{1}{n/k}\sum_{i:(x_i,y_i)\in\mathcal{D}_j}\left(y_i - \hat{f}(x_i;\mathcal{D}\backslash\mathcal{D}_j)\right)^2$$
$$= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i;\mathcal{D}\setminus\mathcal{D}_{j^*(i)})\right)^2$$

where $j^*(i)$ is defined such that $(x_i, y_i) \in \mathcal{D}_{j^*(i)}$.

➔ Note that $E[\mathrm{CV}_{(k)}] = E\big[(Y - \hat{f}(X;\mathcal{D}^{(n-n/k)}))^2\big]$.

Bias-Variance Tradeoff for k-Fold CV (2/2)

- As k increases, the bias of the test error estimate

$$E\big[(Y-\hat{f}(X;\mathcal{D}^{(n-n/k)}))^2\big]-E[(Y-\hat{f}(X;\mathcal{D}))^2]$$

decreases, but the variance

$$E\left[\left\{\frac{1}{n}\sum_{i=1}^{n}\Big[\big(Y_i-\hat{f}(X_i;\mathcal{D}\backslash\mathcal{D}_{j^*(i)})\big)^2-E\big[(Y-\hat{f}(X;\mathcal{D}^{(n-n/k)}))^2\big]\Big]\right\}^2\right]$$

increases.

CV on Classification Problems (1/2)

- Similarly, for classification problems, k-fold CV yields

$$\mathrm{CV}_{(k)}=\frac{1}{k}\sum_{j=1}^{k}\mathrm{Err}_j$$

where $\mathrm{Err}_j=\frac{1}{|\mathcal{D}_j|}\sum_{i\in\mathcal{D}_j}L\big(y_i,\hat{y}_i^{(\mathcal{D}\backslash\mathcal{D}_j)}\big)$.

- In classification problems, we may take

$$L(y,\hat{y})=I\big(y\neq\hat{y}\big)$$

(called the 0-1 loss) or

$$L(y,\hat{f}(x,\mathcal{D}))=-2\log\hat{p}(x;\mathcal{D})=-2\log\hat{\Pr}(y|x;\mathcal{D})$$

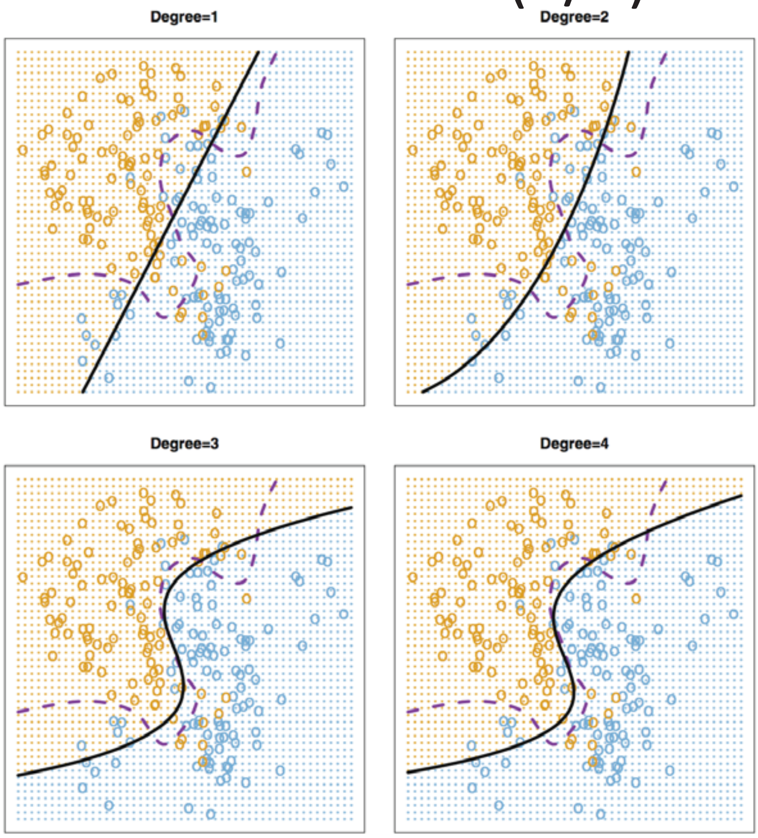
(called the log-likelihood loss, cross-entropy loss, or deviance).

CV on Classification Problems (2/2)

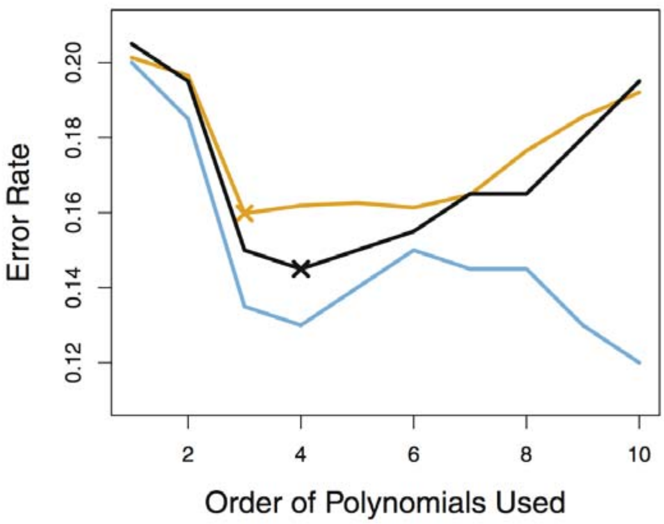
Example:

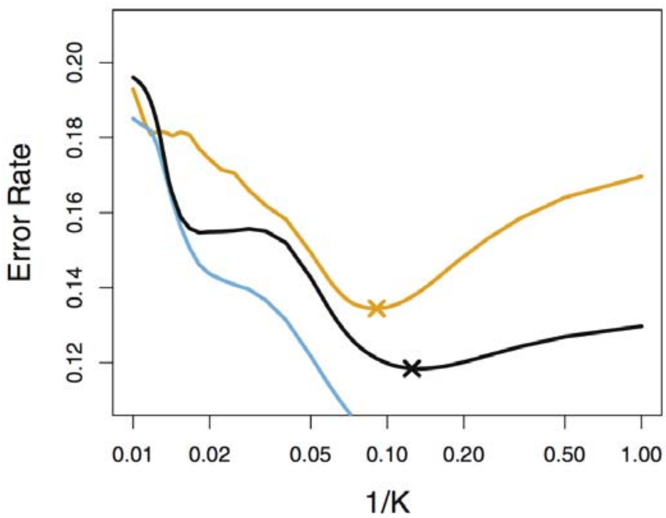
– Polynominal logistic regression

– Test error rates are 0.201, 0.197, 0.160, 0.162 vs Bayes error rate 0.133.



CV for Model Selection





Bootstrap via A Toy Example (1/3)

- Bootstrapping is a tool in statistics (often for measuring accuracy) that involves *random sampling with replacement* of the available data set.

Toy Example:

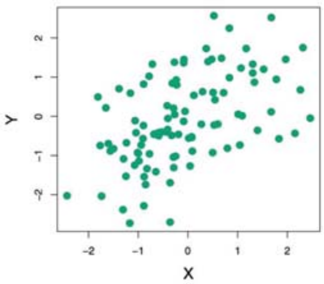
- Investment of a fixed sum of money in two financial assets that yield returns of X and Y, respectively.
- By investing only a fraction α on X and the remaining $1-\alpha$ on Y, the return is $\alpha X + (1-\alpha)Y$.
- The choice of α that minimizes the variability is

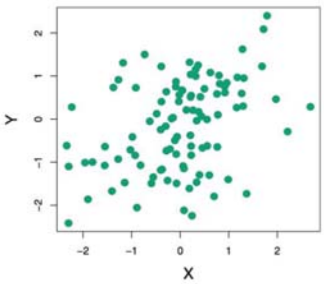
$$\alpha = \arg\min_{\alpha \in [0,1]} \mathrm{Var}(\alpha X + (1-\alpha)Y) = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}.$$

Bootstrap via A Toy Example (2/3)

- Simulate 100 realizations of the values of X and Y using parameters $\sigma_X^2 = 1,\ \sigma_Y^2 = 1.25,\ \sigma_{XY} = 0.5$, and use them to estimate $\sigma_X^2,\ \sigma_Y^2,\ \sigma_{XY}$ and thus α.

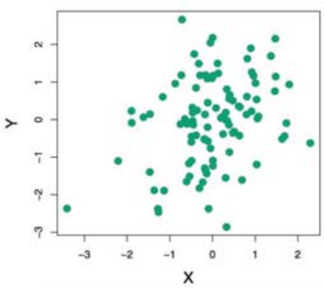
$\hat{\alpha}_1 = 0.576$

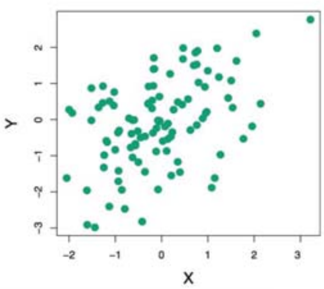




$\hat{\alpha}_2 = 0.532$

$\hat{\alpha}_3 = 0.657$





$\hat{\alpha}_4 = 0.651$

Bootstrap via A Toy Example (3/3)

- By generating 1000 estimates, we can compute the mean $\bar{\alpha} = \frac{1}{1000}\sum_{r=1}^{1000} \hat{\alpha}_r$ and the standard error

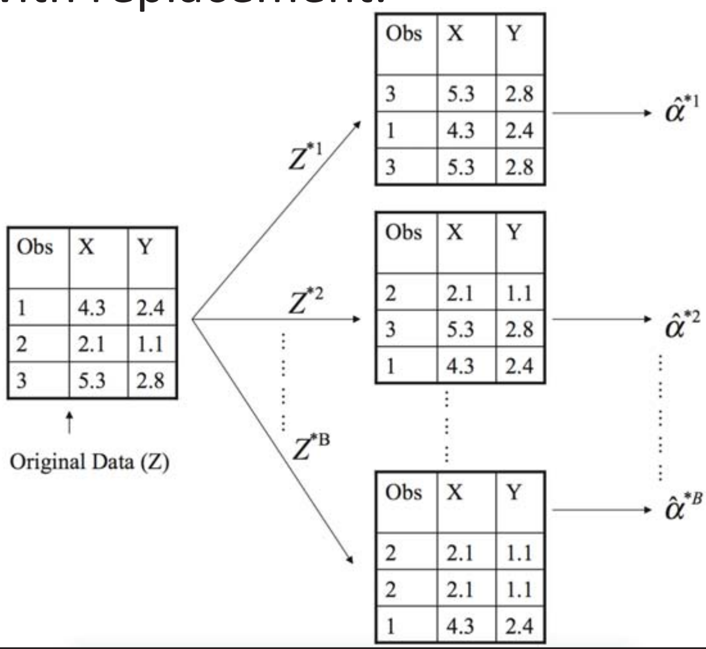
$$\widehat{\mathrm{SE}}(\hat{\alpha}) = \sqrt{\frac{1}{1000-1}\sum_{r=1}^{1000}(\hat{\alpha}_r - \bar{\alpha})^2}.$$

- In practice, we cannot generate data at will to obtain these estimates.

➔**Key idea:** Obtain distinct data sets by repeatedly sampling observations (with replacement) from the original data set. (This is called *bootstrapping*!)

Bootstrap Illustration

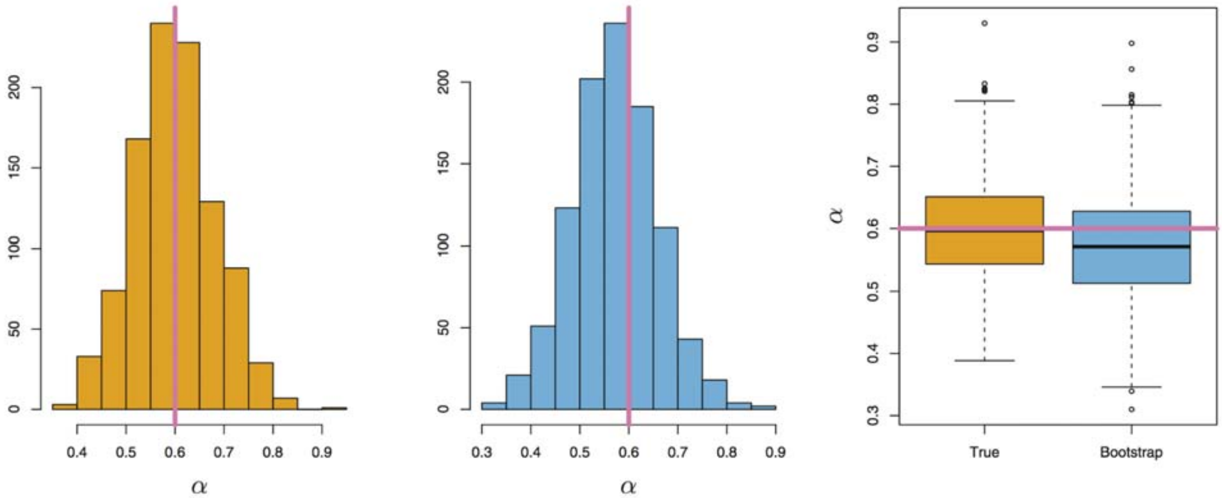
- Generate B bootstrap data sets $Z^{*1}, Z^{*2}, \ldots, Z^{*B}$ by randomly sampling n observations from the original data set Z with replacement.



Bootstrap vs Simulated Data Sets

- The bootstrap estimates are $\hat{\alpha}^{*1}, \hat{\alpha}^{*2}, \ldots, \hat{\alpha}^{*B}$ and the standard error is

$$\widehat{\mathrm{SE}}_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}\left(\hat{\alpha}^{*r} - \frac{1}{B}\sum_{r'=1}^{B}\hat{\alpha}^{*r'}\right)^2}.$$



Bootstrap for Test Error Estimation (1/2)

Question: Can bootstrap be used to estimate test error?

- E.g., use the original data set $\mathcal{D}$ for testing, and bootstrap data sets $\mathcal{D}^{*1}, \mathcal{D}^{*2}, \ldots, \mathcal{D}^{*B}$ for training.
- The estimated test error is

$$\widehat{\mathrm{Err}}_{\mathrm{boot}} = \frac{1}{B}\frac{1}{n}\sum_{b=1}^{B}\sum_{i=1}^{n} L(y_i, \hat{f}(x_i; \mathcal{D}^{*b})).$$

➔ Problem: $\mathcal{D}^{*b}$ contains the (x_i, y_i) with probability

Bootstrap for Test Error Estimation (2/2)

- Let $\mathcal{I}^{(-i)} = \{b : (x_i, y_i) \neq \mathcal{D}^{*b}\}$ be the set of indices of bootstrap data sets that do not include (x_i, y_i). This yields the leave-one-out bootstrap with

$$\widehat{\mathrm{Err}}^{(1)} = \frac{1}{n}\sum_{i=1}^{n}\sum_{b\in\mathcal{I}^{(-i)}} \frac{1}{|\mathcal{I}^{(-i)}|} L(y_i, \hat{f}(x_i; \mathcal{D}^{*b})).$$

➔Problem: The number of distinct observations in each data set is only $0.632n$.

- Adopt a weighted estimate

$$\widehat{\mathrm{Err}}^{(0.632+)} = \left(1 - \frac{0.632}{1-0.368\hat{R}}\right)\overline{\mathrm{err}} + \frac{0.632}{1-0.368\hat{R}}\widehat{\mathrm{Err}}^{(1)}$$

where $\hat{R} = \frac{\widehat{\mathrm{Err}}^{(1)} - \overline{\mathrm{err}}}{\frac{1}{n^2}\sum_i \sum_{i'} L(y_i, \hat{f}(x_{i'}; \mathcal{D})) - \overline{\mathrm{err}}}.$