Data Science II
(P8106)

Department of Biostatistics
Mailman School of Public Health
Columbia University

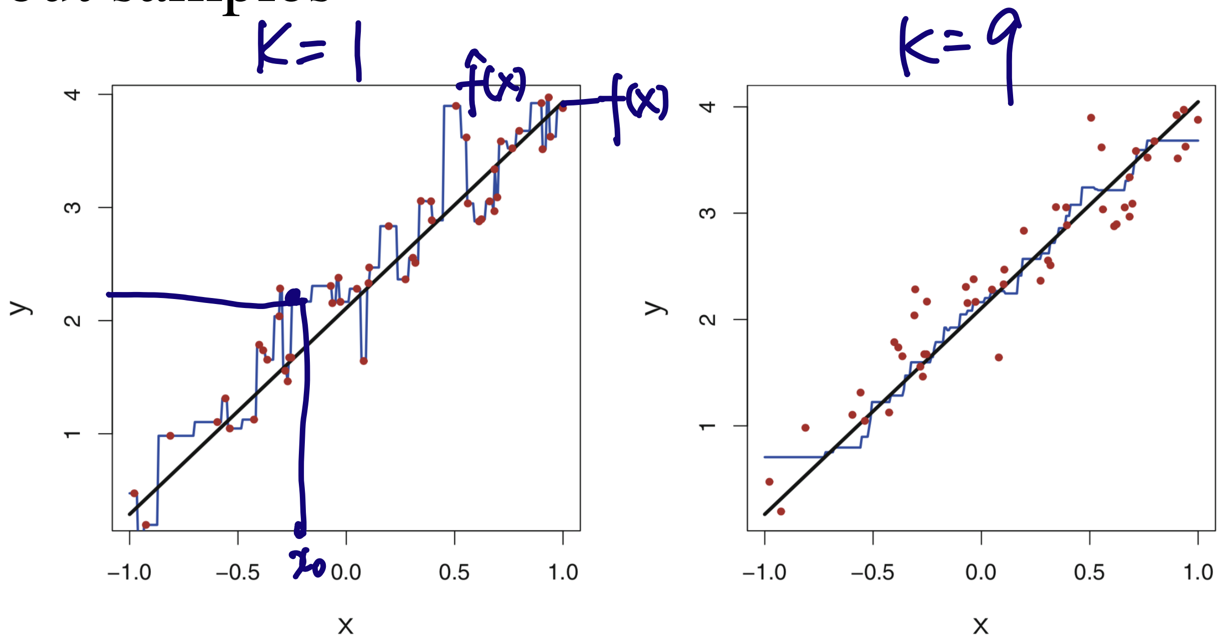
Spring 2024

Resampling Methods

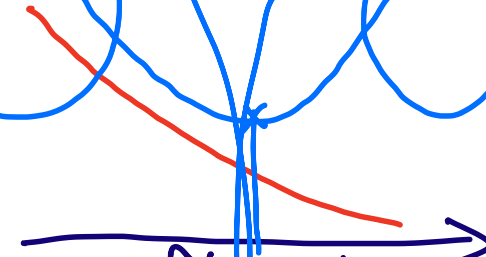
- Refit a model of interest to samples formed from the training data to obtain additional information about the fitted model
- **Cross-validation:** estimate the prediction error
- **Bootstrap:** evaluating the variance of an estimator

Model tuning

- Many models have important parameters which cannot be directly estimated from the data
- Example: k-nearest neighbor
- Tuning parameter - no analytical form to calculate an appropriate value
- Given a candidate set of tuning parameters, one can determine the optimal value based on the performance on hold-out samples



Training error vs. test error



- Test error is the average error that results from using a statistical learning method to predict the response on a new observation
- Training error is calculated by applying the statistical learning method to the observations used in its training
- The training error can dramatically underestimate the test error

How to estimate test error?

- Best solution: a large designed test set, often not available!
- Cross-validation: hold out a subset of the training data from the fitting process

Training error vs. test error

- Test error is the average error that results from using a statistical learning method to predict the response on a new observation
- Training error is calculated by applying the statistical learning method to the observations used in its training
- The training error can dramatically underestimate the test error

How to estimate test error?

- Best solution: a large designed test set, often not available!
- Cross-validation: hold out a subset of the training data from the fitting process

AIC, BIC

error/loss + penalty

Validation-set approach

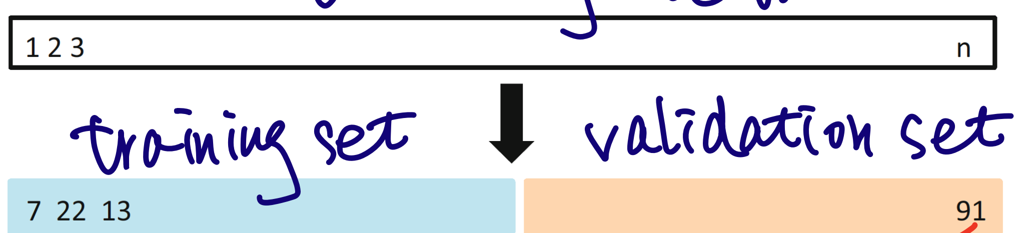
training data

- **Randomly** divide the available set of samples into two parts: training set and validation set
- The model is fit on the training set
- The fitted model is used to predict the responses for the observations in the validation set
- The resulting validation-set error provides an estimate of the test error (e.g., MSE)

data

test data

training ~~data~~ data



$\hat{f}$

[ISL] Figure 5.1

$$MSE = \frac{\sum_{i \in VS} (y_i - \hat{f}(x_i))^2}{n_{VS}}$$

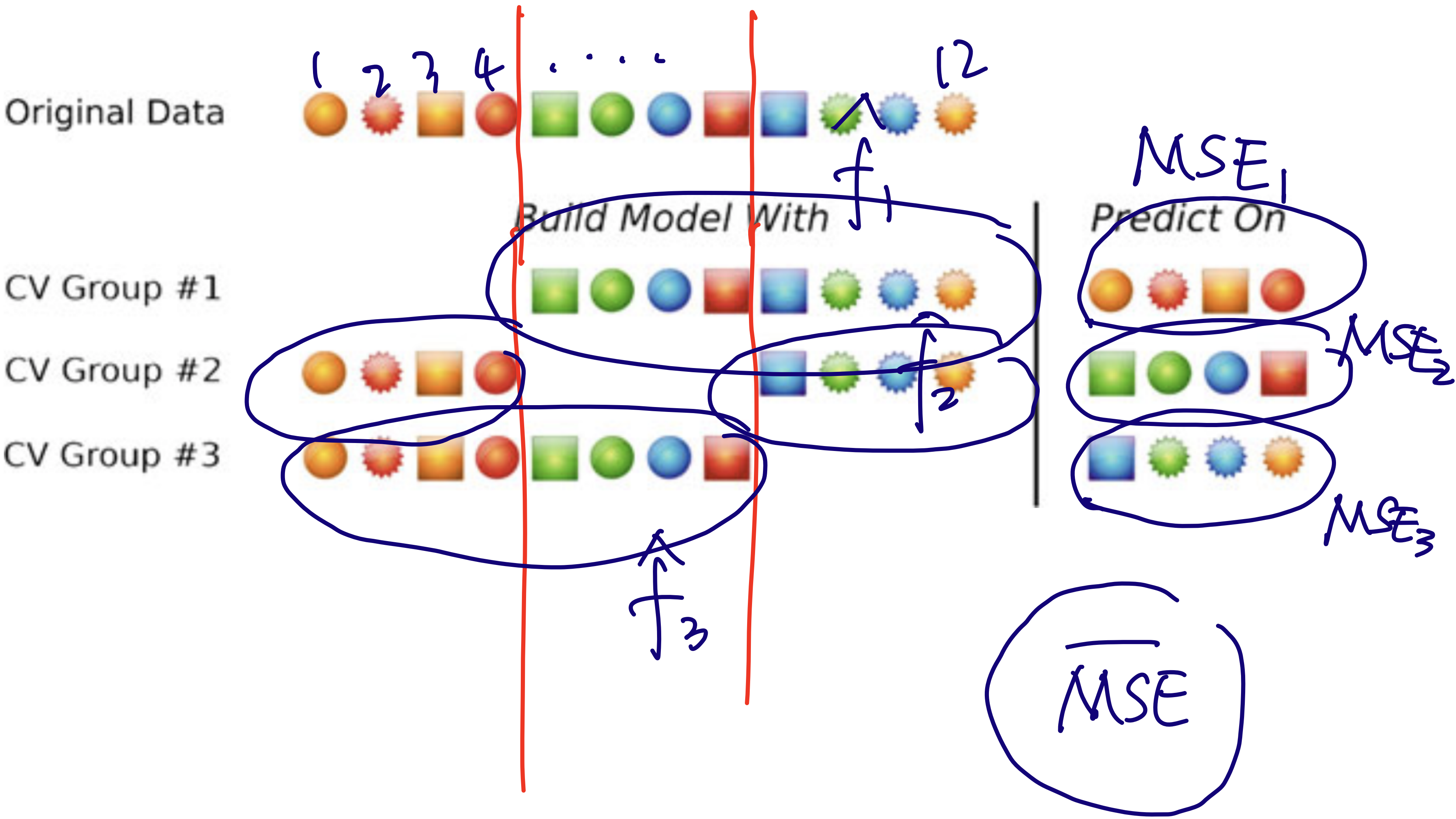
Drawbacks of validation set approach

- The validation estimate of the test error can be highly variable
- Only a subset of observations are used to fit the model
- This suggests that the validations set error may tend to overestimate the test error for the model fit on the entire set

K-fold cross-validation

- Widely used approach for estimating test error
- Estimates can be used to select best model, and to give an idea of the test error of the final chosen model
- Idea
  - Randomly divide the data into K equal-sized parts
  - Leave out part k, fit the model to the other $K - 1$ parts (combined)
  - Obtain predictions for the left-out kth part
  - This is done in turn for each part $k = 1, 2, \ldots, K$, and then the results are combined

The details: three-fold CV



The details

- Let the K parts be $C_1, C_2, \ldots, C_K$ where C_k denotes the indices of the observations in part k
- There are n_k observations in part k: if n is a multiple of K, then $n_k = n/K$
- Compute

CV MSE

$$CV_{(K)} = \sum_{k=1}^{K} \frac{n_k}{n} MSE_k$$

  - $MSE_k = \sum_{i \in C_k} (y_i - \hat{y}_i)^2 / n_k$
  - $\hat{y}_i$ is the fit for observation i, obtained from the data with C_k removed
  - Since each training set is $(K-1)/K$ as big as the original training data, the estimates of prediction error will typically be **biased upward**

A nice special case

- Setting $K = n$ yields n-fold or *leave-one-out cross-validation* (LOOCV)
- In least-squares linear regression, the following formula holds

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

$x_i\hat{\beta}$

  - $\hat{y}_i$ is the ith fitted value from the original least squares fit
  - h_i is the leverage (diagonal of the "hat" matrix)

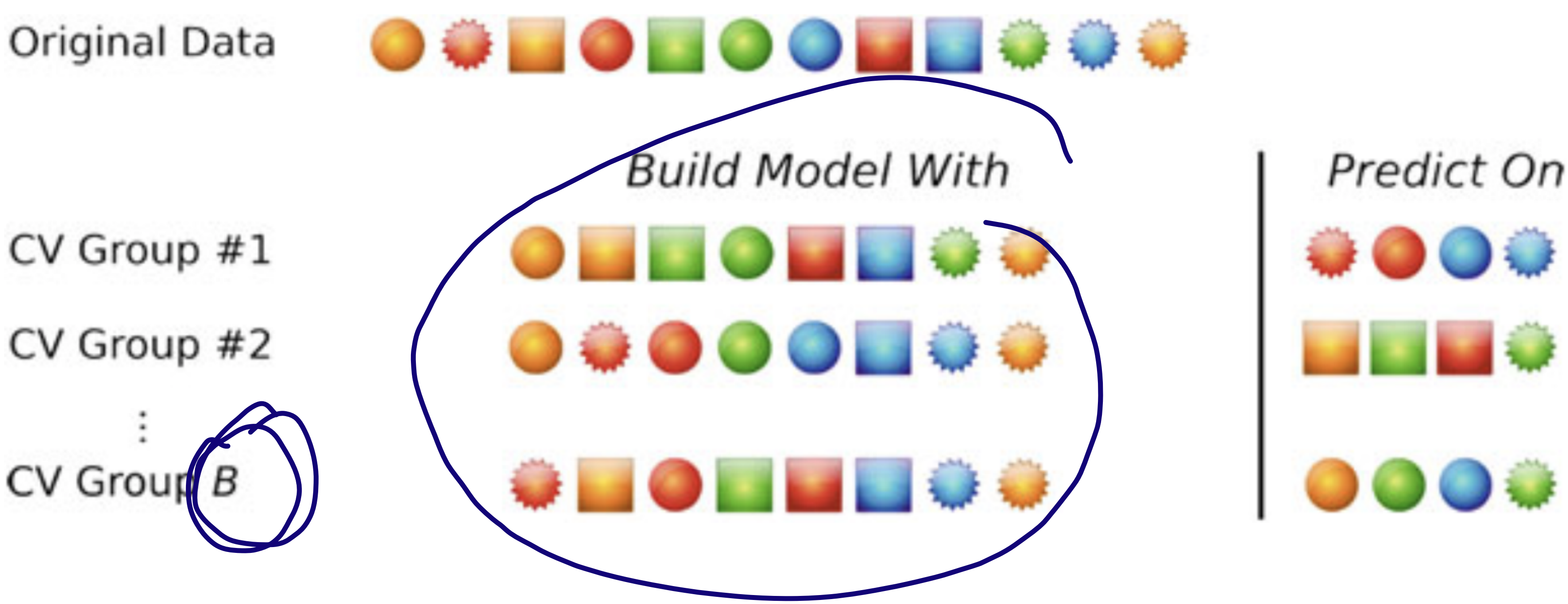
LOOCV

$k=10$

- Advantages
  - **Less bias**
  - No randomness in the training/validation set splits
- LOOCV is sometimes useful, but typically doesn't *shake up* the data enough
- The estimates from each fold are highly correlated and hence their average can have **high variance**
- Bias-variance tradeoff

Monte Carlo cross-validation

- Also known as "leave-group-out cross-validation"
- Rule-of-thumb proportion of the modeling set: 75%~80%
- Number of repetitions is larger (e.g., 50-200)
- Increasing B decreases the uncertainty of the performance estimates

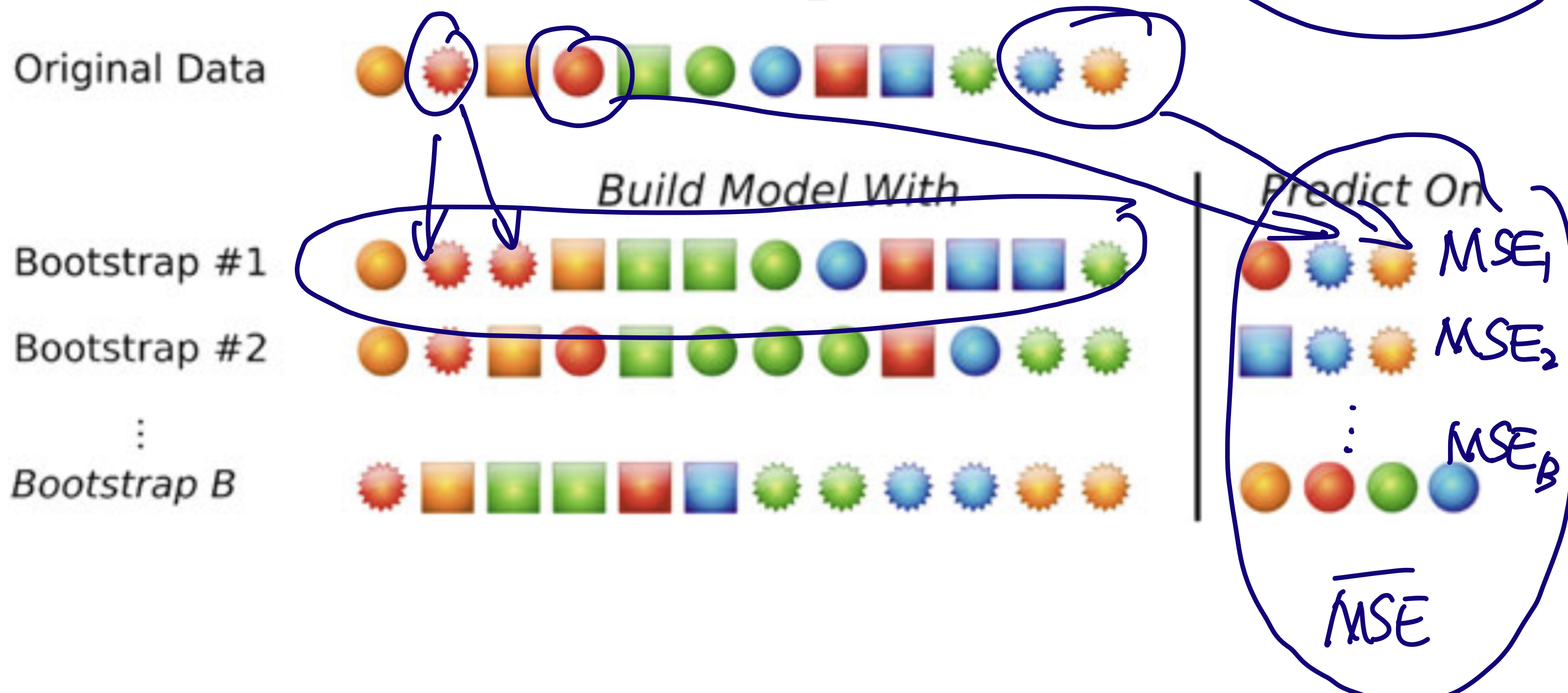


The 632 bootstrap

$P(i \in \text{bootstrap sample}) \approx 0.632$

$= 1 - \left(\frac{n-1}{n}\right)^n \xrightarrow{n \to \infty} 1 - 1/e$ ✓

- Sampling with **replacement**, **same size** as the original sample
- On average, each bootstrapping sample contains $0.632n$ unique observations
- The 632 method: $0.632 \times$ bootstrap error $+ \; 0.368 \times$ training error



Cross-validation: right and wrong

$\mathrm{Cor}(Y, X_j)$

1. Starting with 5000 predictors, find the 100 predictors having the largest correlation with the response
2. We then fit a model using only these 100 predictors

How do we estimate test performance using CV?

A. Apply cross-validation in step 2

B. Apply cross-validation in steps 1 and 2