Resampling: Cross Validation and Bootstrap

(ISLR 5.1 - 5.2)

Yingbo Li

Southern Methodist University

STAT 4399

Outline

Resampling methods

How:

1. Repeatedly drawing samples from a training set, and
2. Refitting a model of interest on each sample.

What:

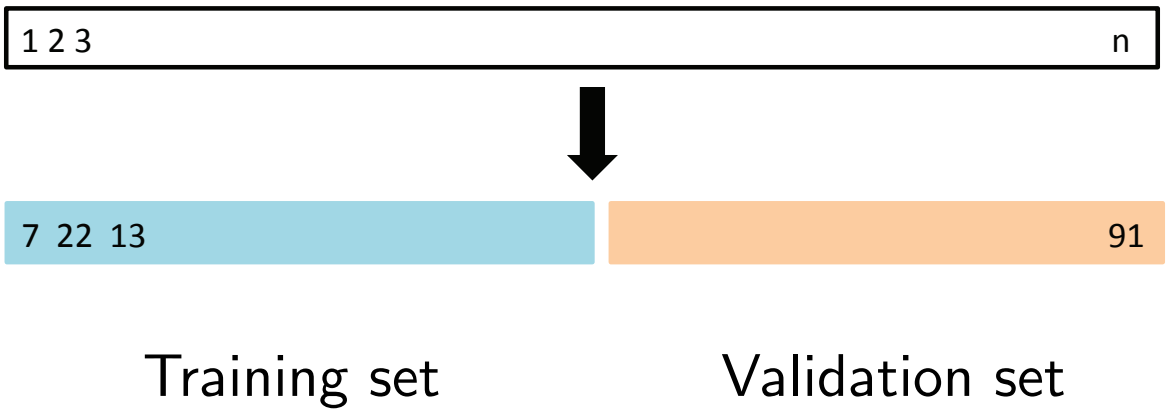
- *Cross validation*: hold out a subset of the training set as the test set
- *Bootstrap*: generate a new dataset from the current dataset by sampling with replacement

Why:

- Model assessment: estimate test error rates (CV)
- Model selection: select the appropriate level of model flexibility (CV)
- Estimate standard errors of point estimators (bootstrap)

The validation set approach

1. Randomly divide the current data in two halves: a training set and a validation set (or hold-out set)
2. Fit the model using the training set
3. Based on the fitted mode, predicted using the validation set, report
   - MSE (regression), or
   - misclassification rate (classification)



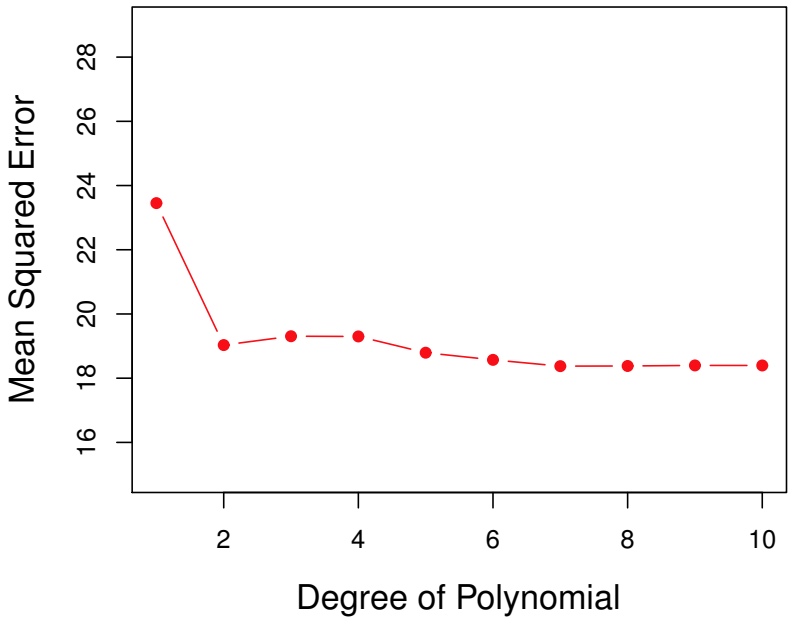
Example: `Auto` data

- Study the relationship between `mpg` and `horsepower`
- Compare linear vs higher-order polynomial terms in a linear regression

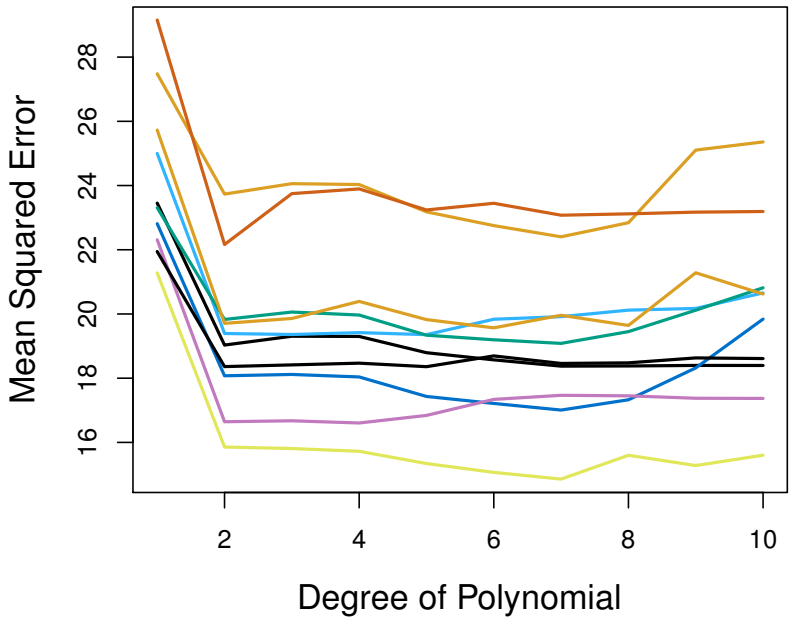
$$\texttt{mpg} \sim \texttt{horsepower}$$
$$\texttt{mpg} \sim \texttt{horsepower} + \texttt{horsepower}^2$$
$$\texttt{mpg} \sim \texttt{horsepower} + \texttt{horsepower}^2 + \texttt{horsepower}^3$$
$$\vdots$$

- We randomly split the 392 observations into two sets,
  - a training set containing 196 of the data points, and
  - a validation set containing the remaining 196 observations.
- Fit all candidate models using the training data
- Calculate predict MSEs using the validation data.
- The model with the lowest validation (testing) MSE is the winner!

A single split



Repeated 10 times



- The quadratic model yields much smaller MSE than the linear model.
- A lot of variability among different splits.
- In general, models perform worse when training on fewer observations, so test error may be overestimated.

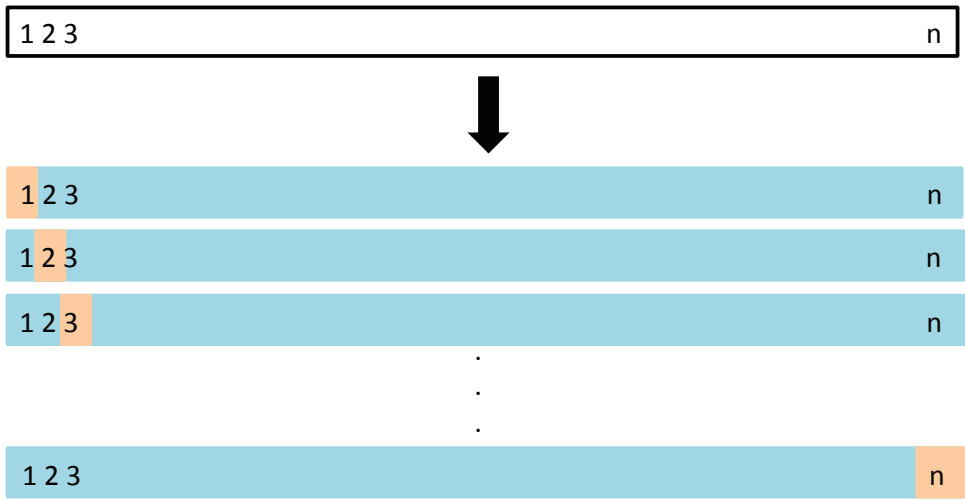
Leave one out cross validation (LOOCV)

For each $i = 1, 2, \ldots, n$,

- Hold out a single data point observation (Y_i, X_i)
- Use the rest $n - 1$ observations as the training data to fit the model
- Predict the hold-out point and compute $\text{MSE}_i = (Y_i - \hat{Y}_i)^2$

- Estimation of test error

$$\text{CV}_{(n)} = \frac{1}{n} \sum_{i=1}^{n} \text{MSE}_i$$

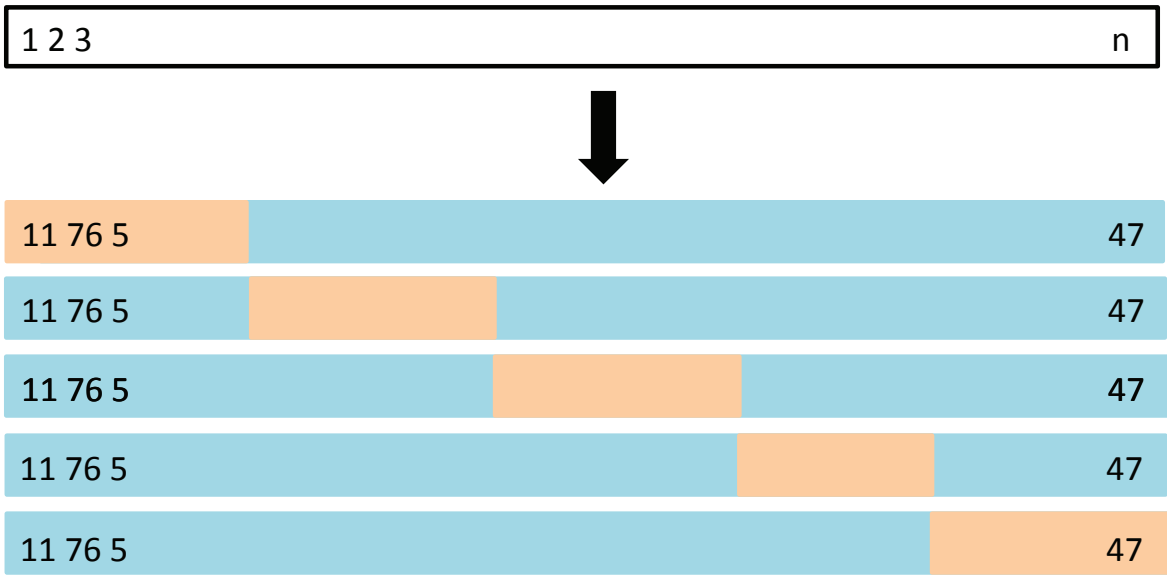


LOOCV vs. the validation set approach

- LOOCV's advantages
  - Almost all the data set is used in the training sets.
  - No randomness in the training / validation set split.
- LOOCV's disadvantages
  - Computationally intensive: fit the model n times!

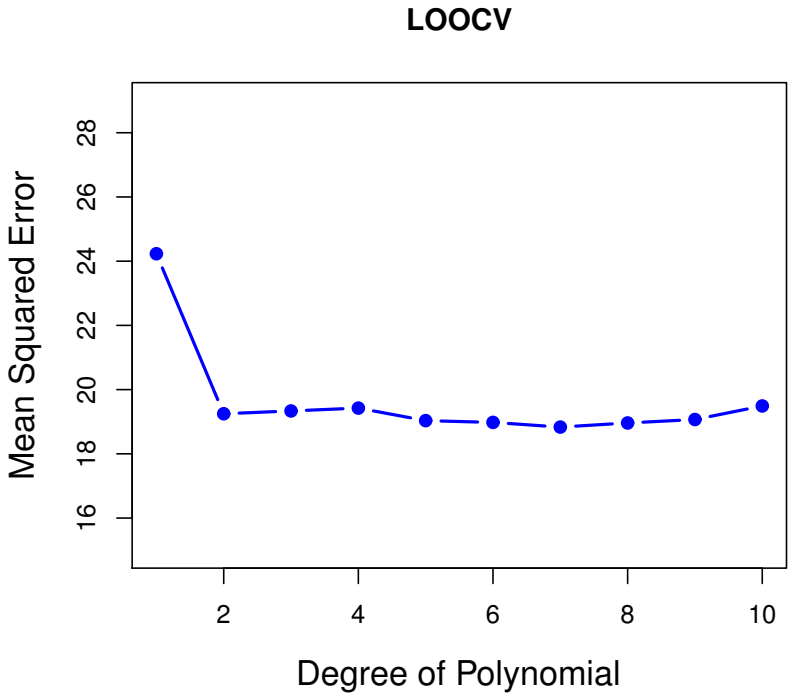
k-fold cross validation

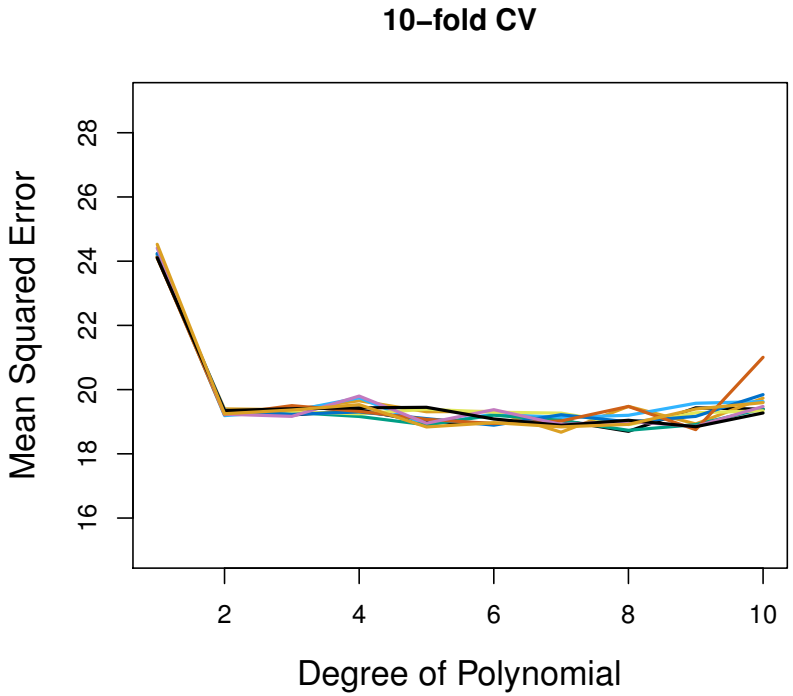
- Divide the data into k folds (typically, $k = 5$ or 10)
- Hold out the first fold, fit the model on the remaining $k - 1$ folds, and see how good the predictions are on the left out (i.e. compute the MSE on the first part)
- Repeat this for k times, holding out a different fold each time
- Estimate the test MSE using the average of the k different MSE's



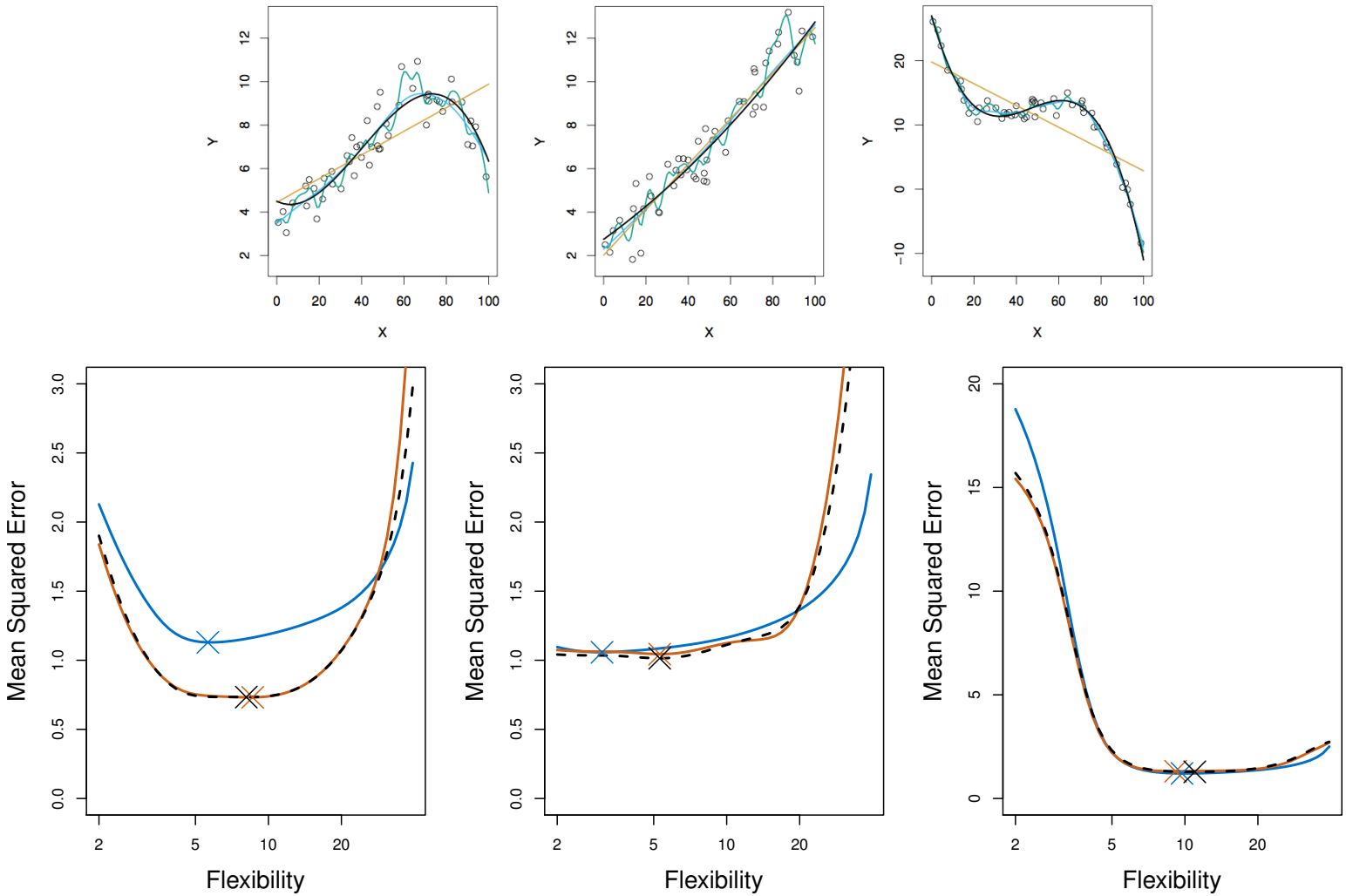
Auto data: mpg $\sim$ horsepower, polynomial regression

- Since each training set is only $\frac{k-1}{k}$ as big as the original training set, the estimates of prediction error will typically be biased upward.
- LOOCV is a special case of CV: $k = n$.
- Both are stable, LOOCV is more computational intensive.
- Variability of 10-fold CV is lower than the validation set approach.





10-fold CV on three simulated data



True test MSE, 10-fold CV MSE, LOOCV MSE.

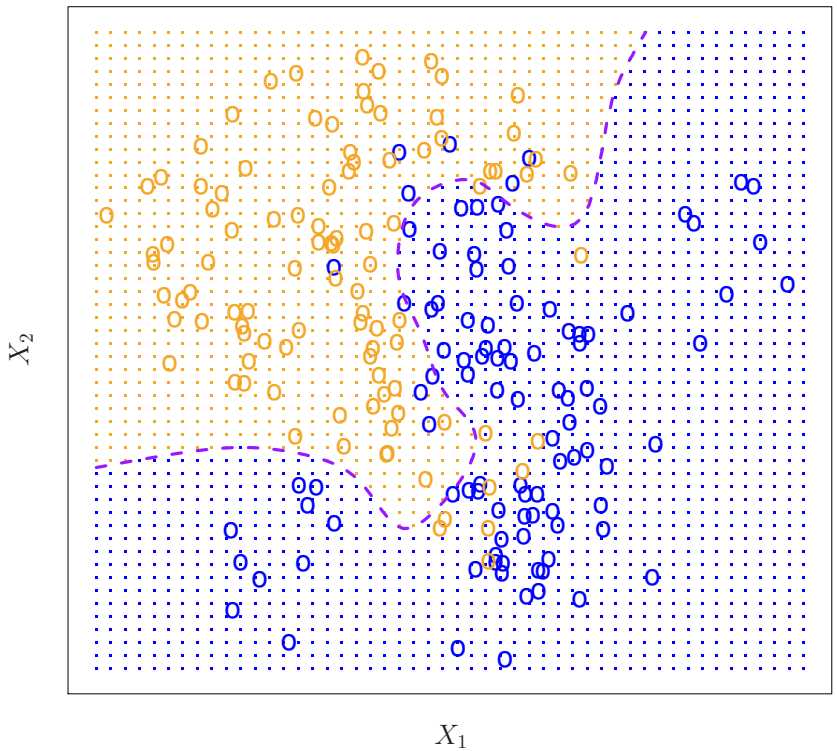
Bias-variance trade-off of k-fold CV

- Putting aside that LOOCV is more computationally intensive than k-fold CV (when $k < n$)
- LOOCV is less bias than k-fold CV ...
  - Since each training set is smaller than the original training set, the estimates of prediction error will typically be biased upward.
- But, LOOCV has higher variance than k-fold CV
  - LOOCV doesn't shake up the data enough. The estimates from each fold are highly correlated, so their average can have high variance.

CV on classification problems

Average the misclassification (error) rates instead of the MSE's.

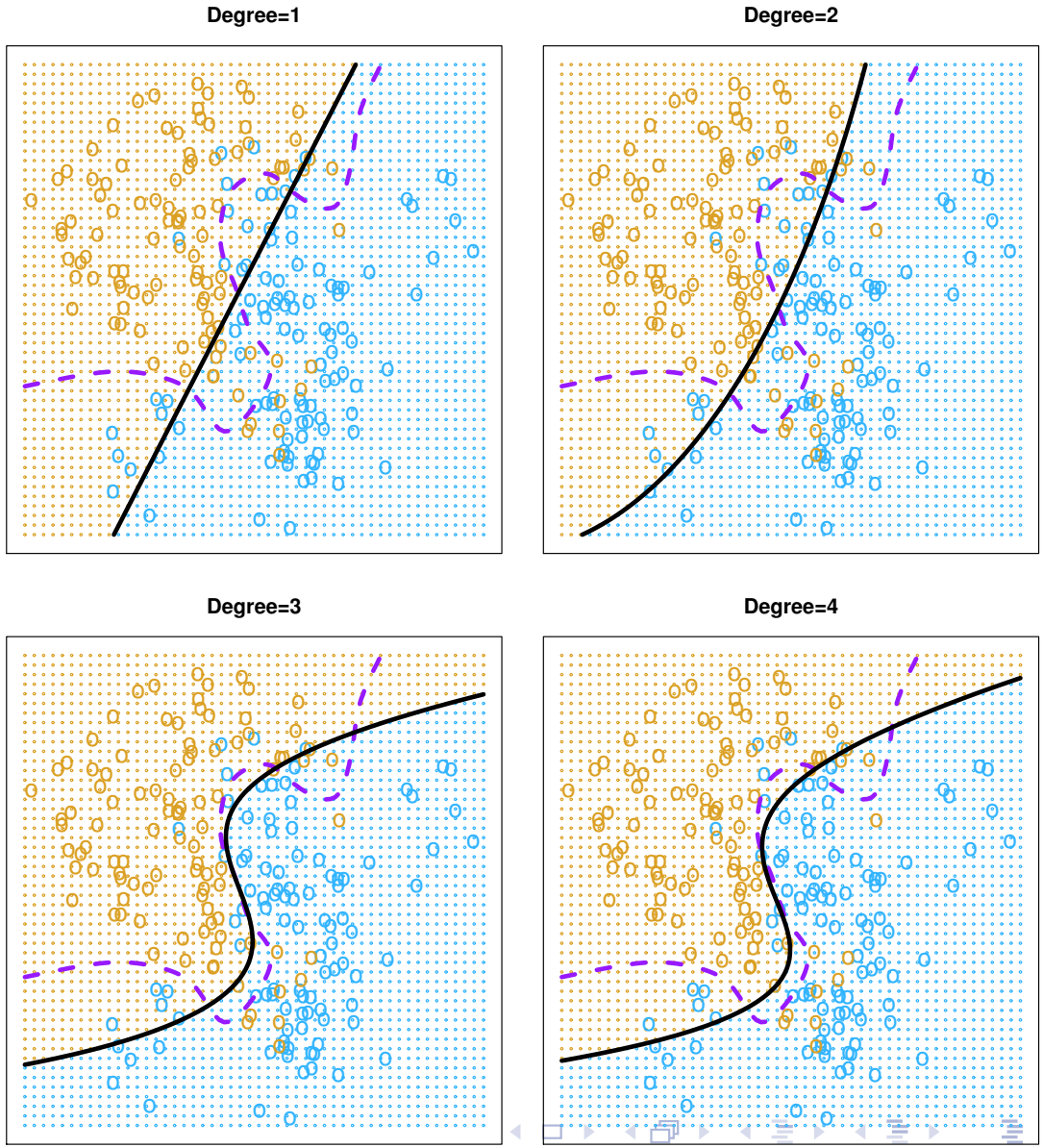
A classification example: $p = 2$



Bayes' error rate: 0.133

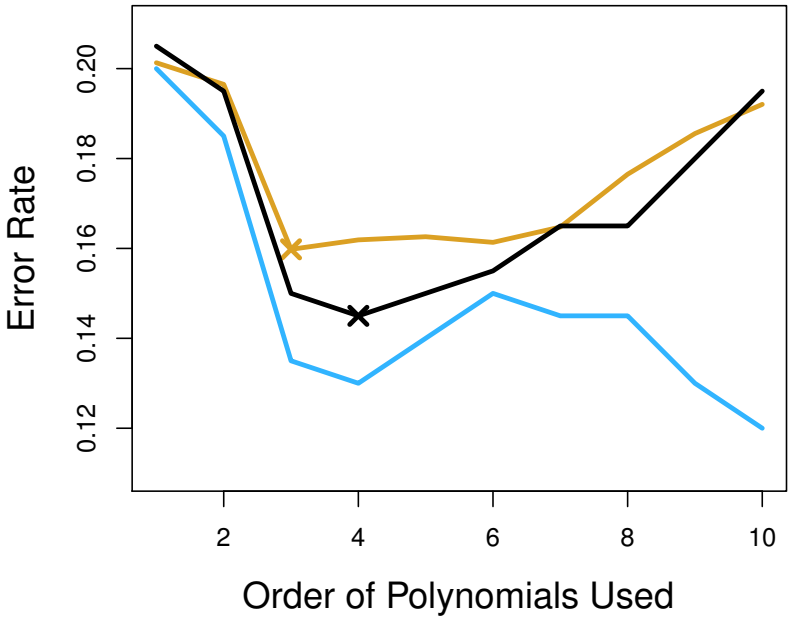
Polynomial logistic regressions

Order	Test error rate
1	0.201
2	0.197
3	0.160
4	0.162
Bayes	0.133

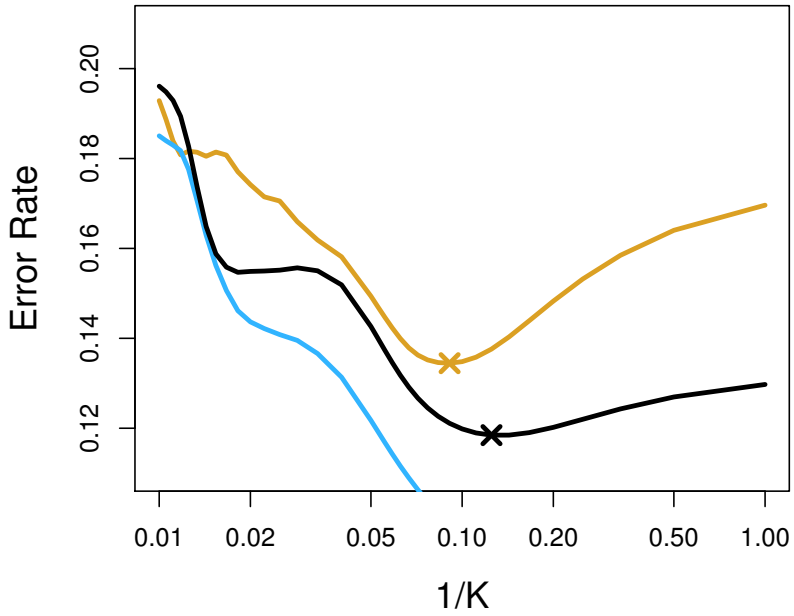


Example: use CV to do model selection

Polynomial logistic regression



KNN



Training error, Test error, 10-fold CV error.

The bootstrap

- The *bootstrap* is a flexible and powerful statistical tool that can be used to *quantify the uncertainty* associated with a given estimator or statistical learning method.
  - standard error of a parameter
  - confidence interval for a parameter
- The use of the term bootstrap derives from the phrase

  to pull oneself up by one's bootstraps,

  thought to be based on one of the eighteenth century "The Surprising Adventures of Baron Munchausen" by Rudolph Erich Raspe: *The Baron had fallen to the bottom of a deep lake. Just when it looked like all was lost, he thought to pick himself up by his own bootstraps.*

Example: the standard error doesn't have a closed form

- Suppose that we wish to invest a fixed sum of money in two financial assets that yield returns of X and Y, respectively, where X and Y are random quantities.
- We will invest a fraction α of our money in X, and will invest the remaining $1-\alpha$ in Y.
- We wish to choose α to minimize the total risk, or variance, of our investment.

$$\hat{\alpha} = \arg\min_{\alpha} \text{Var}\left(\alpha X + (1-\alpha)Y\right).$$

- One can show that the $\hat{\alpha}$ has a closed form

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

where $\hat{\sigma}_X^2, \hat{\sigma}_Y^2$ are sample variances and $\hat{\sigma}_{XY}$ is the sample covariance.

- How to estimate the variability of the estimation $SE(\hat{\alpha})$?

Bootstrap samples

In a perfect world, suppose we know the true population:

1. Simulate a dataset of n paired observations (X, Y), and compute $\hat{\alpha}$
2. Repeat Step 1 for 1000 times, so we obtain 1000 estimates for α:

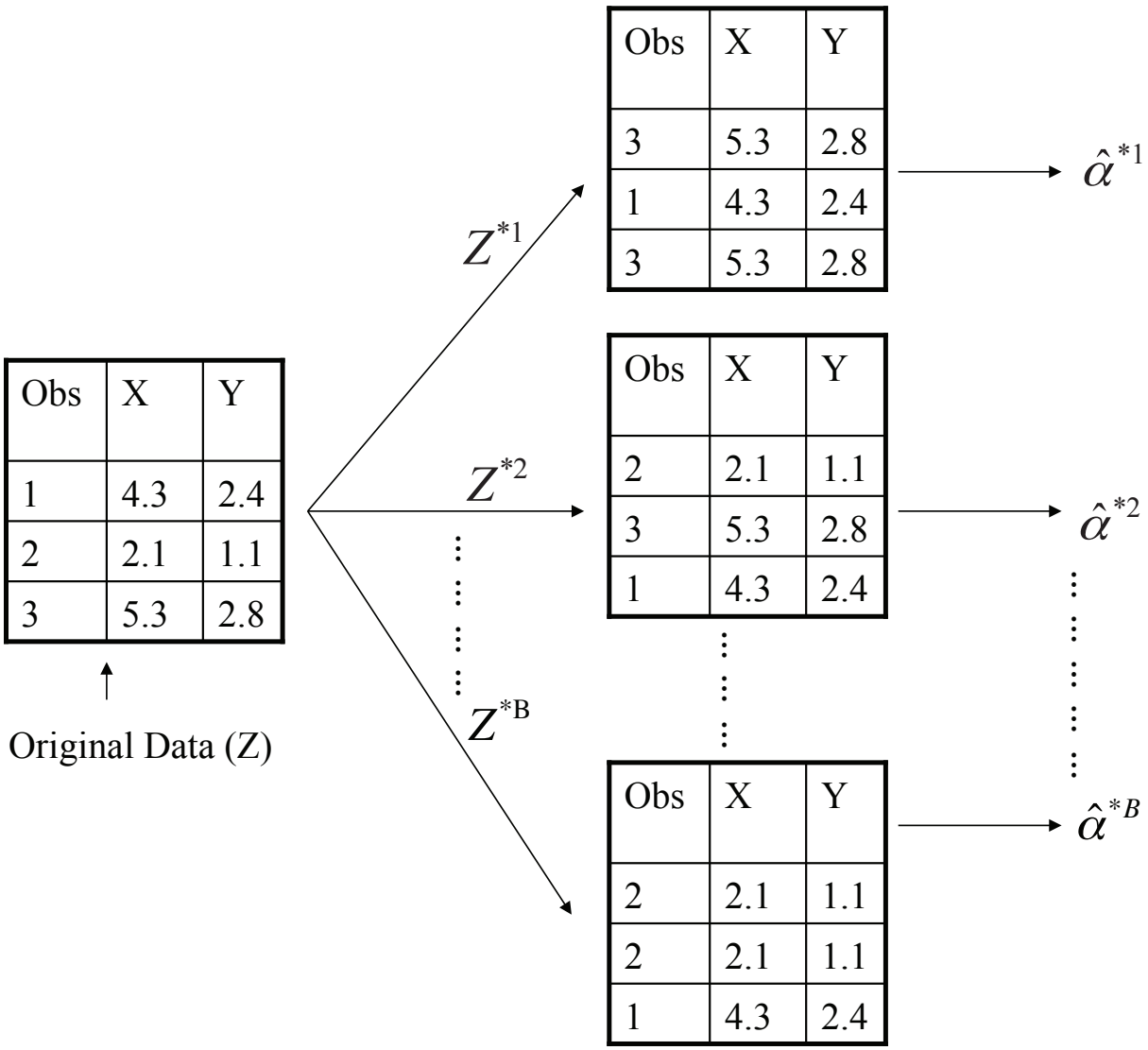
$$\hat{\alpha}_1, \hat{\alpha}_2, \cdots, \hat{\alpha}_{1000}$$

3. Compute the sample standard deviation of them as $SE(\hat{\alpha})$.

In the real world, we don't know the true distributions, and only have one dataset of n observations.

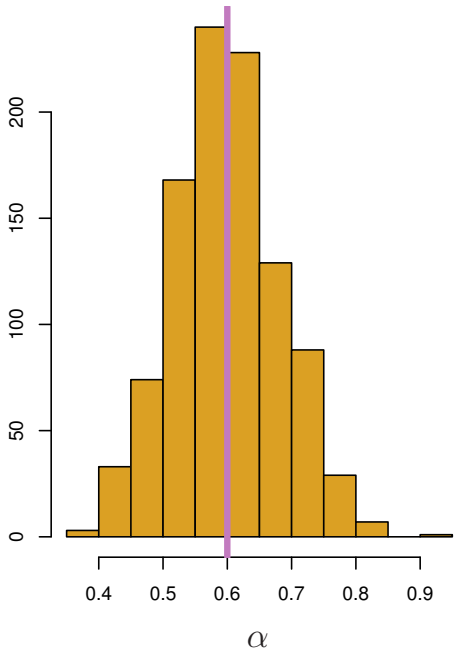
- We just need to modify above Step 1:
  Draw n observations from the original data set *with replacement*.
- As a result, some observations may appear more than once in a given bootstrap data set, and some not at all.

A bootstrap sampling illustration of $n = 3$

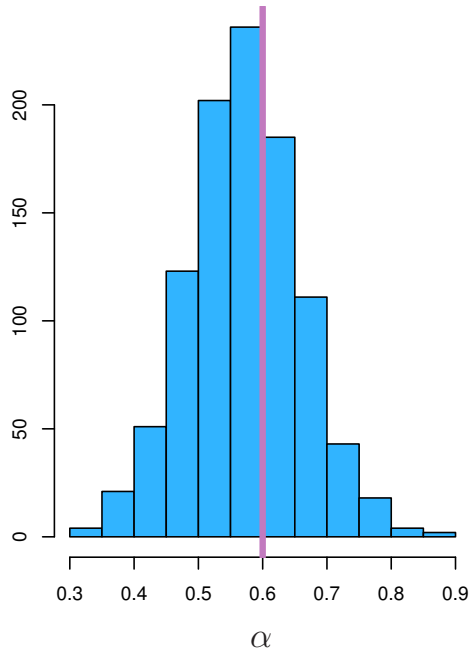


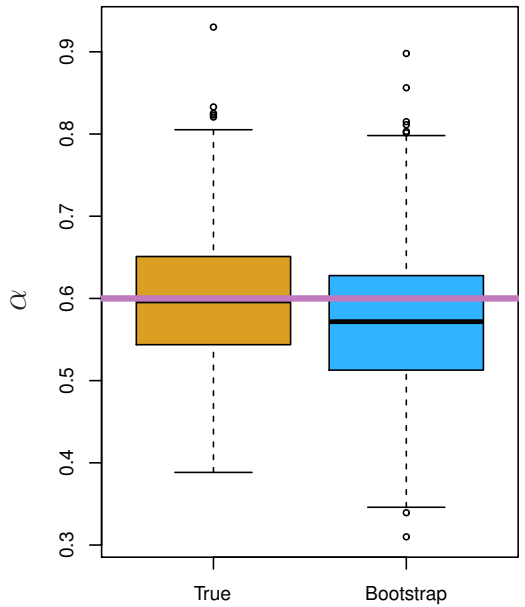
1000 samples, each of size $n = 100$.

Independent samples



Bootstrap samples





How to estimate 95% confidence interval?

Bootstrap percentile confidence interval: report the 2.5% and 97.5% percentile of $\{\hat{\alpha}_1, \hat{\alpha}_2, \cdots, \hat{\alpha}_{1000}\}$.