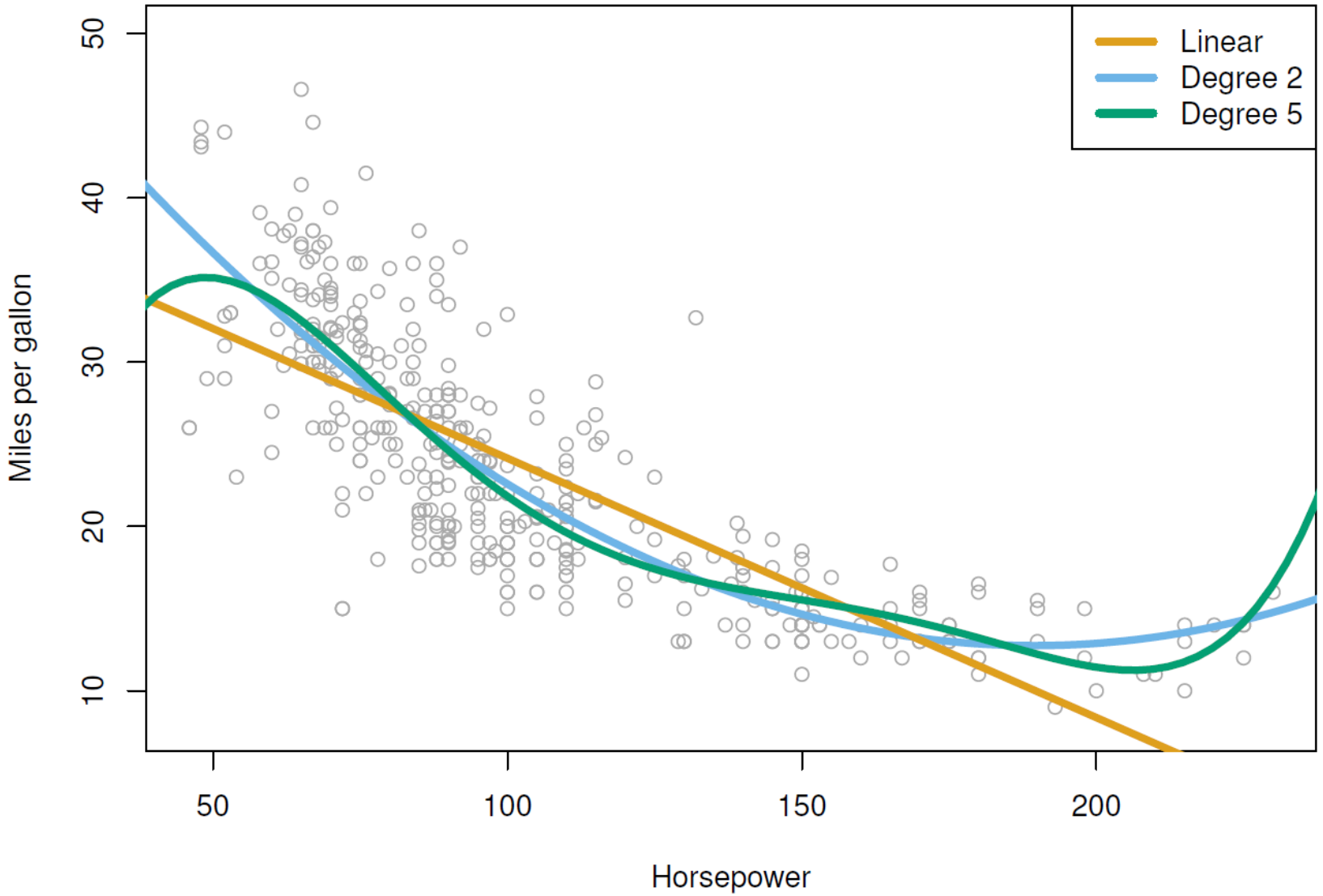
RESAMPLING METHODS

Outline

- Cross-validation
- The bootstrap

- How could we estimate the test error associated with a statistical learning method?
- How could we measure the accuracy of a parameter estimate or of a statistical learning method?



Some of the figures and tables in this presentation are taken from "*An Introduction to Statistical Learning, with Applications in R*" (Springer) with permission from the authors: G. James, D. Witten, T. Hastie, and R. Tibshirani

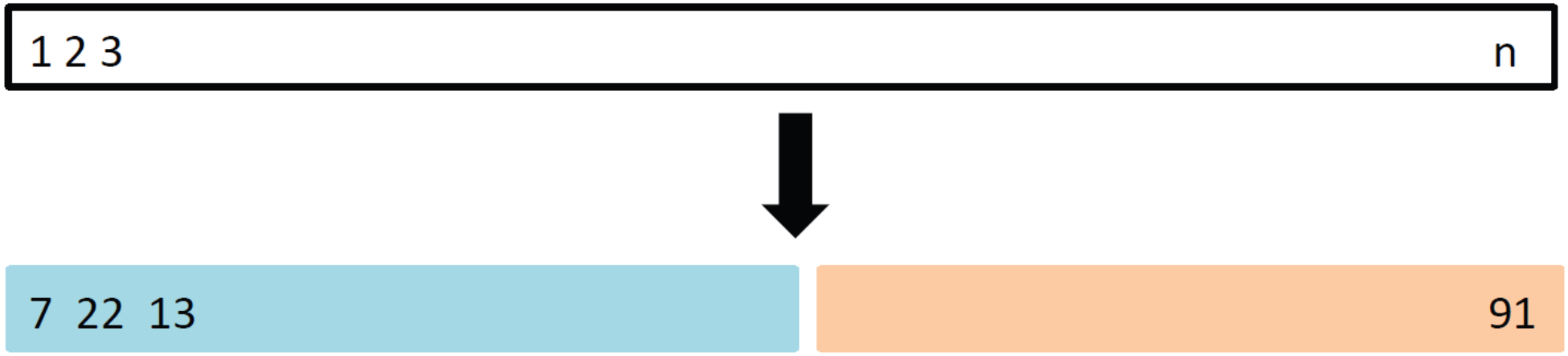
Auto data

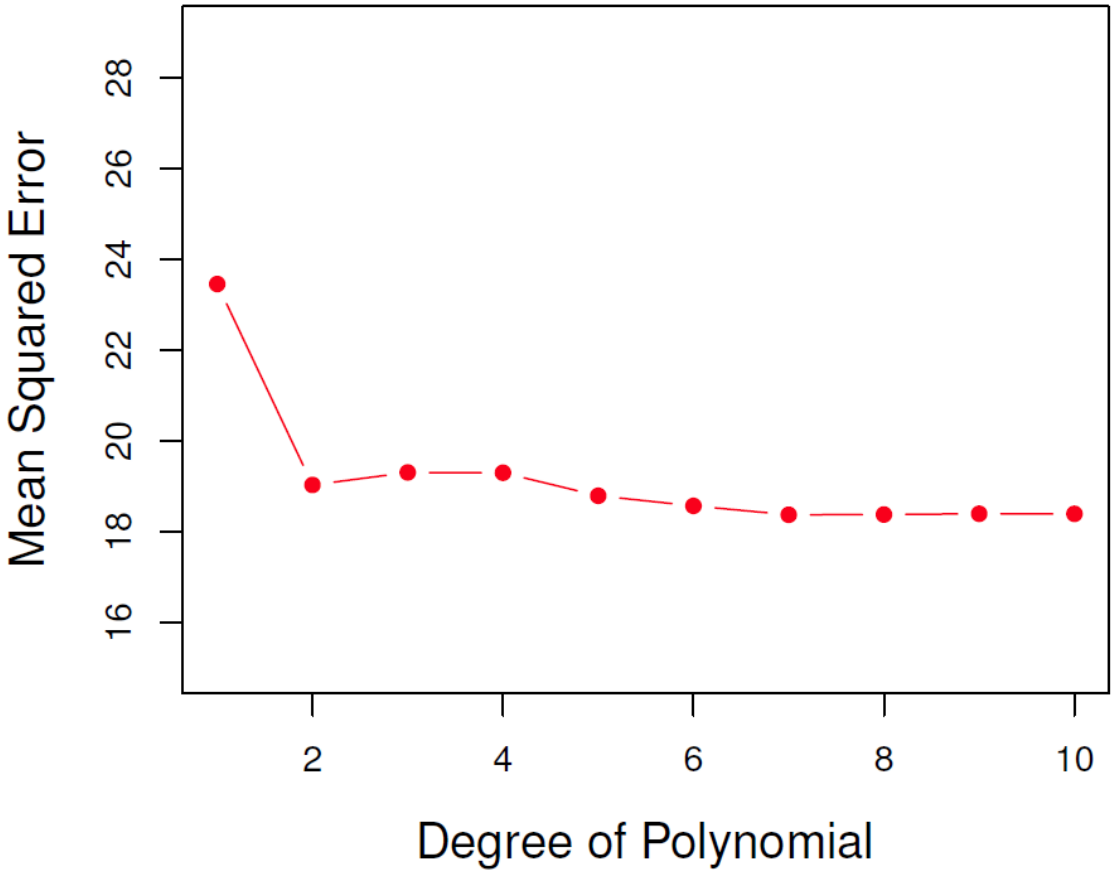
- A non-linear relationship between mpg and horsepower
  - A linear model that predicts mpg using horsepower and horsepower^2 gives better results
- A cubic or higher-order fit?

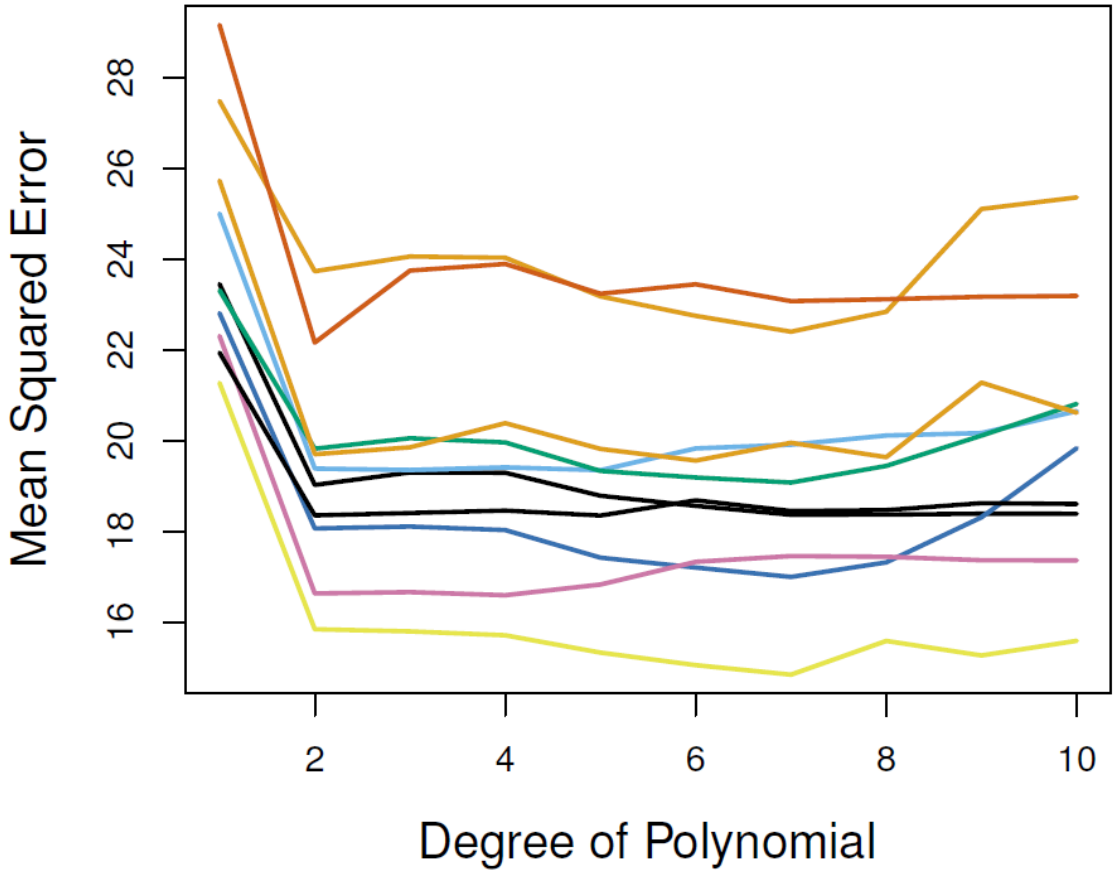
Cross-validation

The validation set approach

- A set of n observations are randomly split into a training set and a validation set
- The statistical learning method is fit on the training set, and its performance is evaluated on the validation set







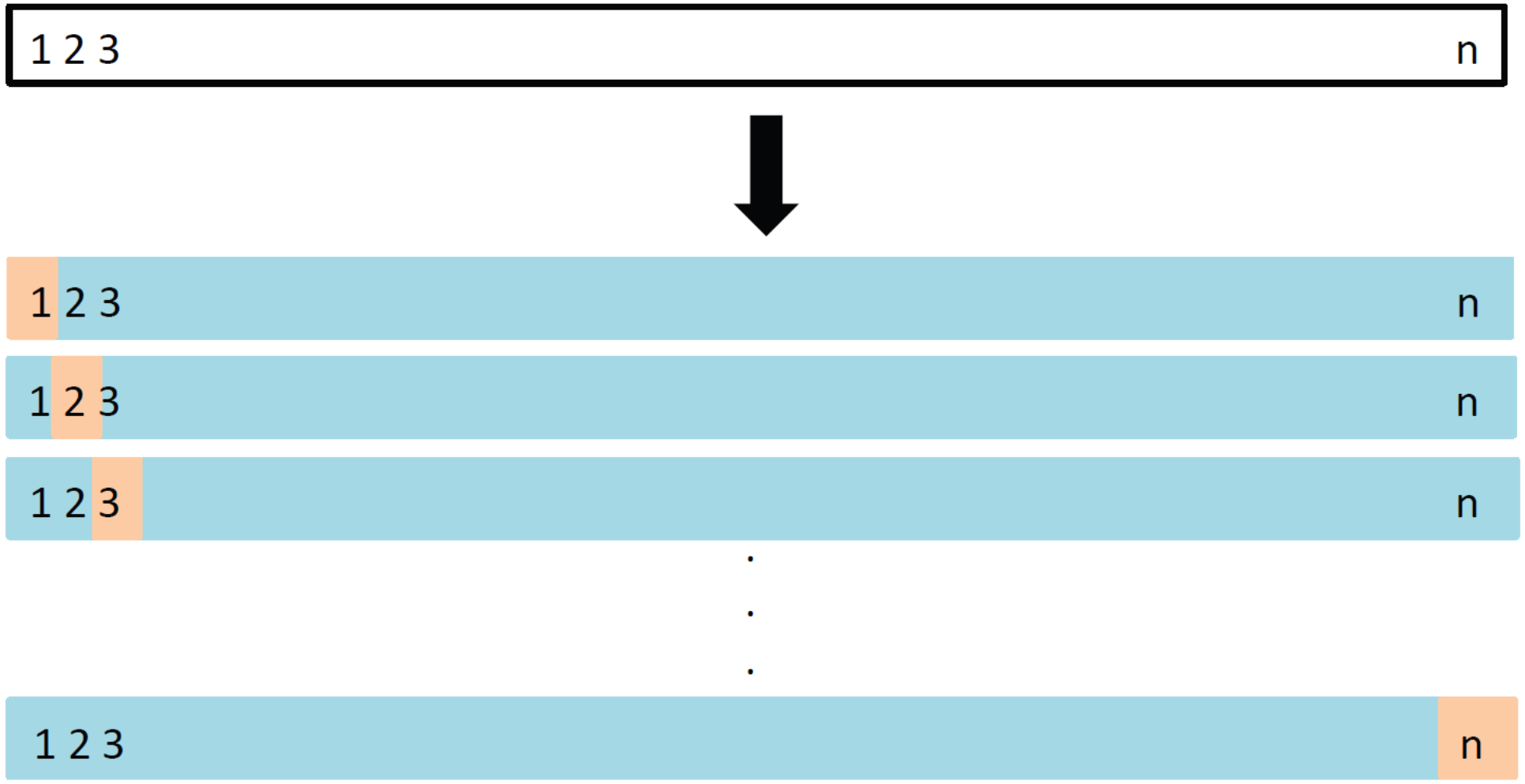
Auto data. The validation set approach was used to estimate the test error that results from predicting mpg using polynomial functions of horsepower

Potential drawbacks

- The validation estimate of the test error rate can be highly variable
- The validation set error rate may tend to overestimate the test error rate for the model fit on the entire data set

Leave-one-out cross-validation

- A set of n data points is repeatedly split into a training set containing all but one observation, and a validation set that contains only that observation
  - The first training set contains all but observation 1, the second training set contains all but observation 2, …



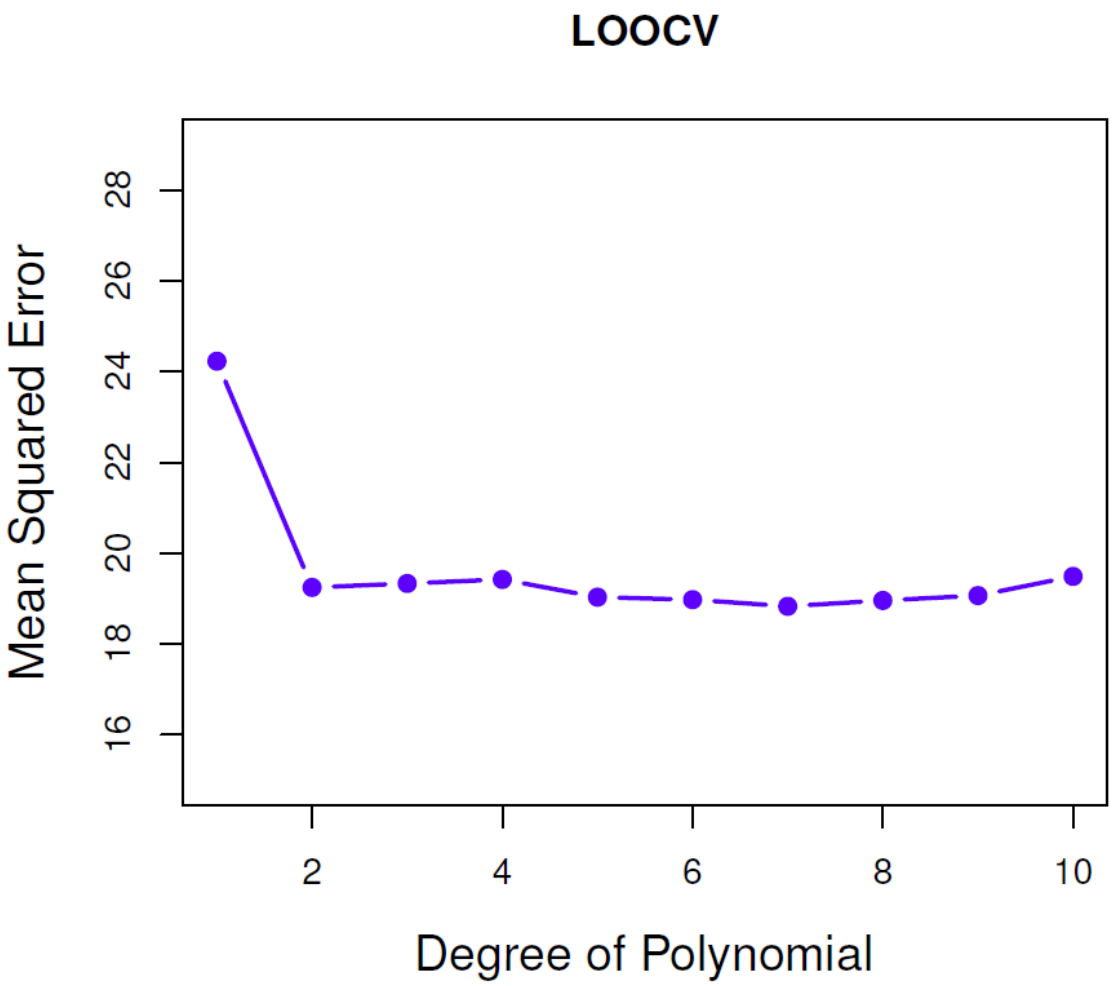
A schematic display of LOOCV

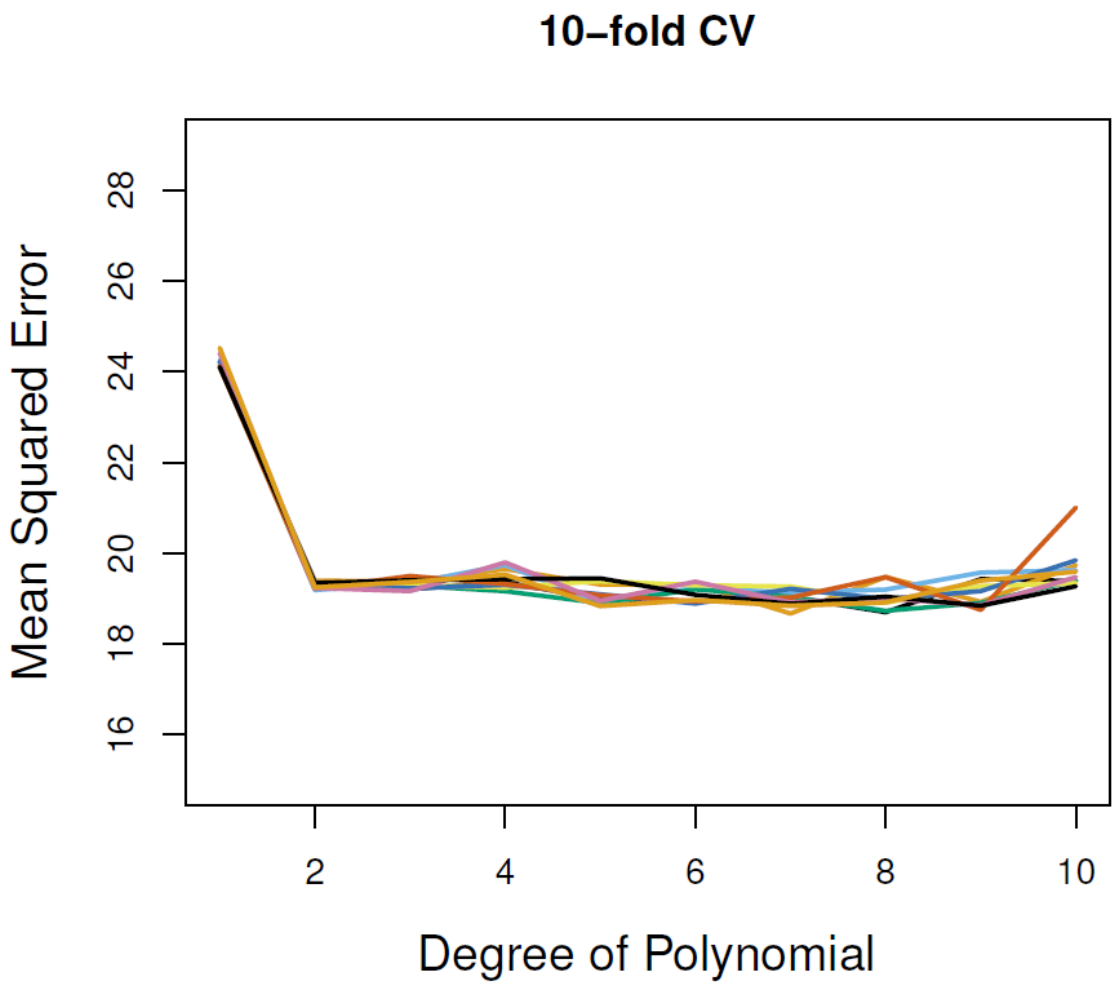
- The statistical learning method is fit on the training observations, and a prediction is made for the excluded observation
  - $\hat{y}_1, \hat{y}_2, \ldots$
  - $MSE_1 = (y_1 - \hat{y}_1)^2, MSE_2 = (y_2 - \hat{y}_2)^2, \ldots$

The LOOCV error

- The test error is estimated by averaging the n MSEs

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





Auto data. Cross-validation was used to estimate the test error that results from predicting mpg using polynomial functions of horsepower

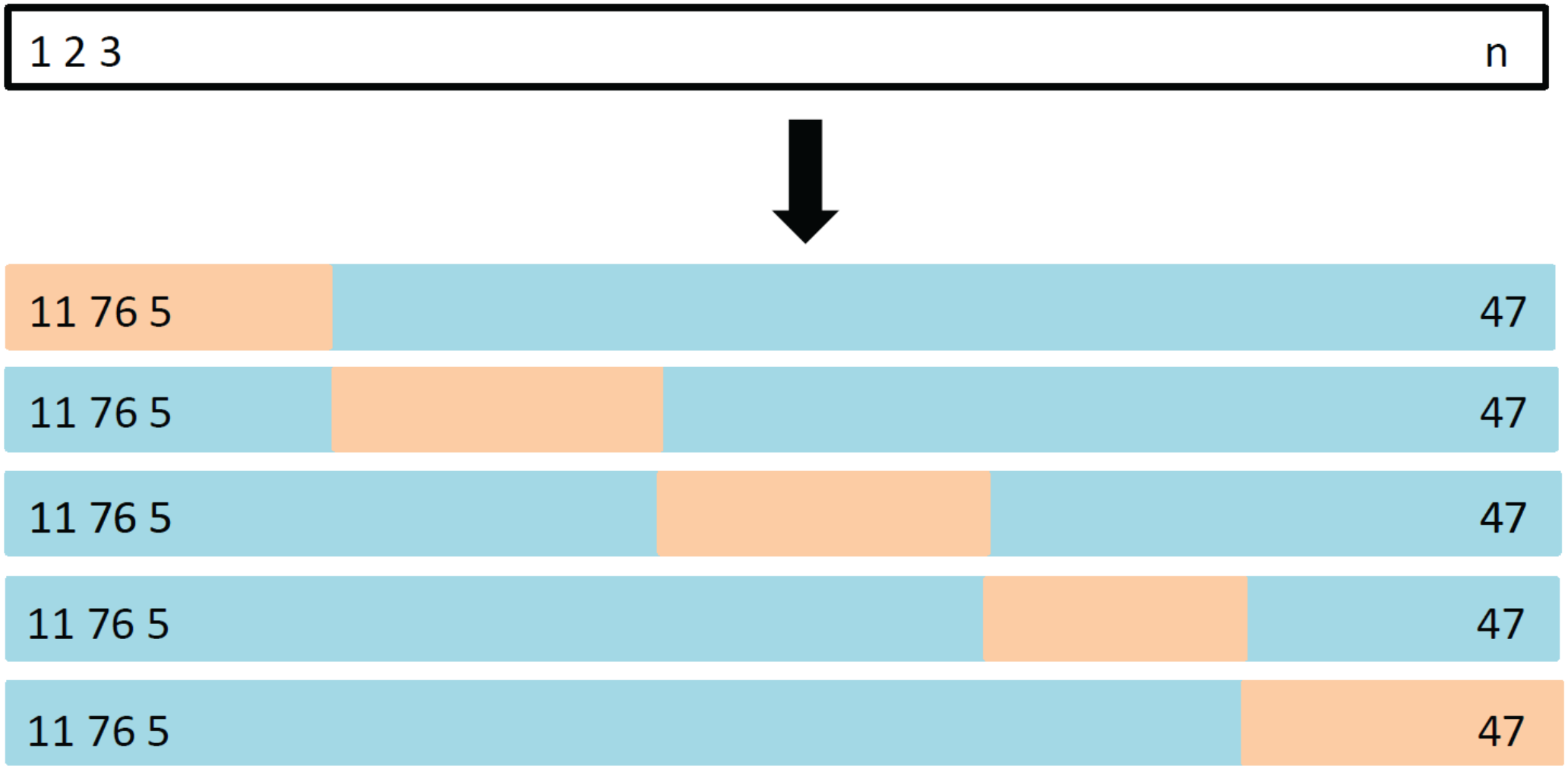
Properties of LOOCV

- Performing LOOCV multiple times will always yield the same results
  - No randomness in the training/validation set splits
- LOOLV has far less bias
  - It tends not to overestimate the test error rate as much as the validation set approach does

- LOOCV requires fitting the statistical learning method n times
  - It has the potential to be computationally expensive

k-fold cross-validation

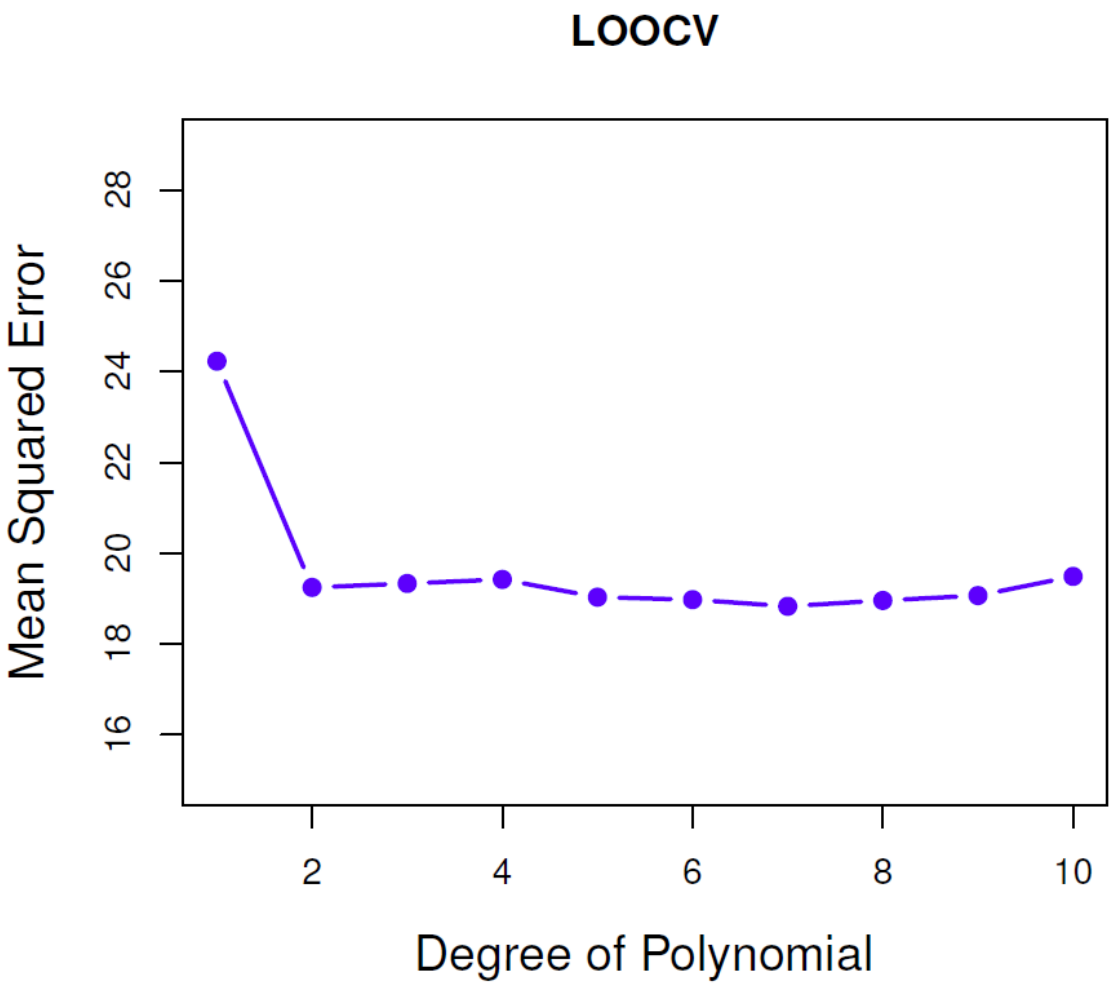
- A set of n observations is randomly split into k groups, or folds
- For each $i \in \{1, 2, \ldots, k\}$, the ith fold is treated as a validation set

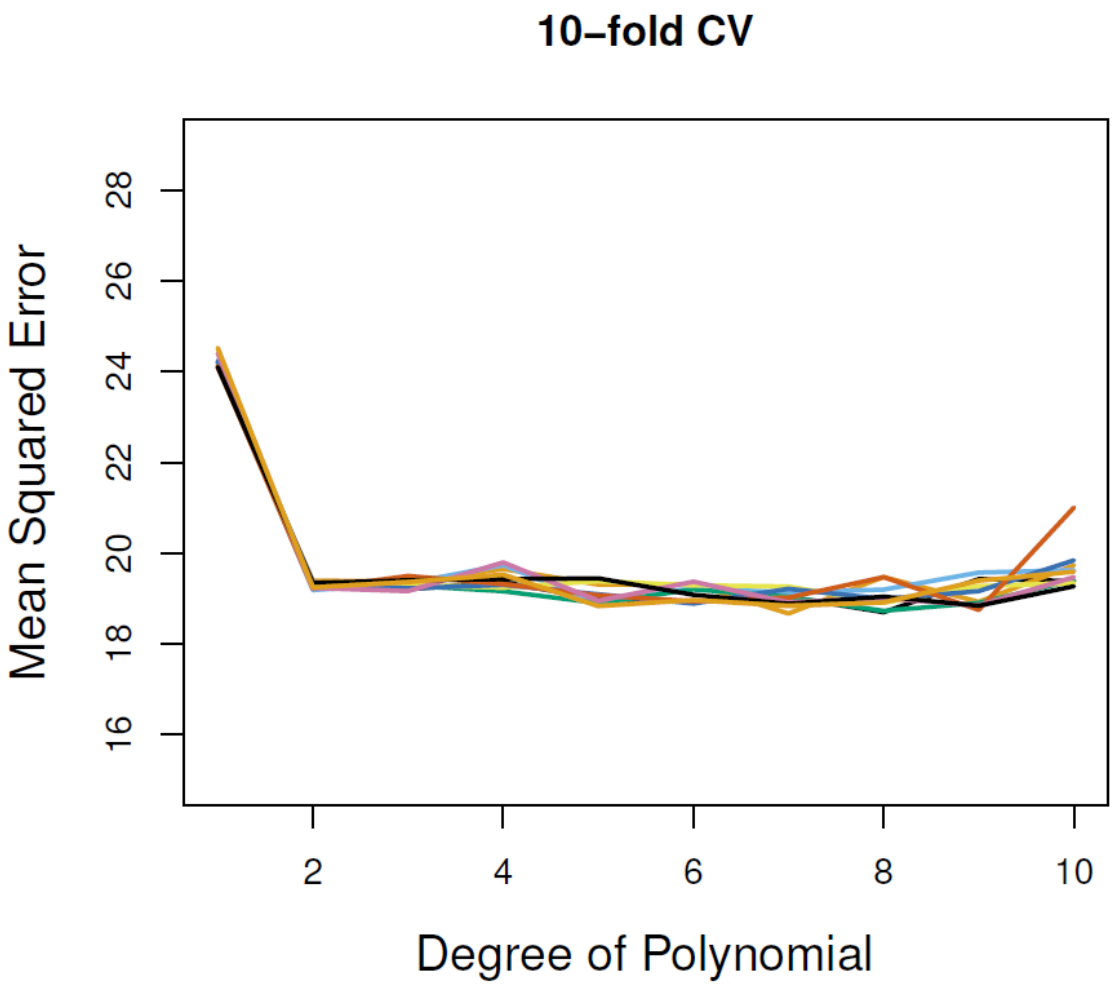


A schematic display of 5-fold CV

- The test error is estimated by

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

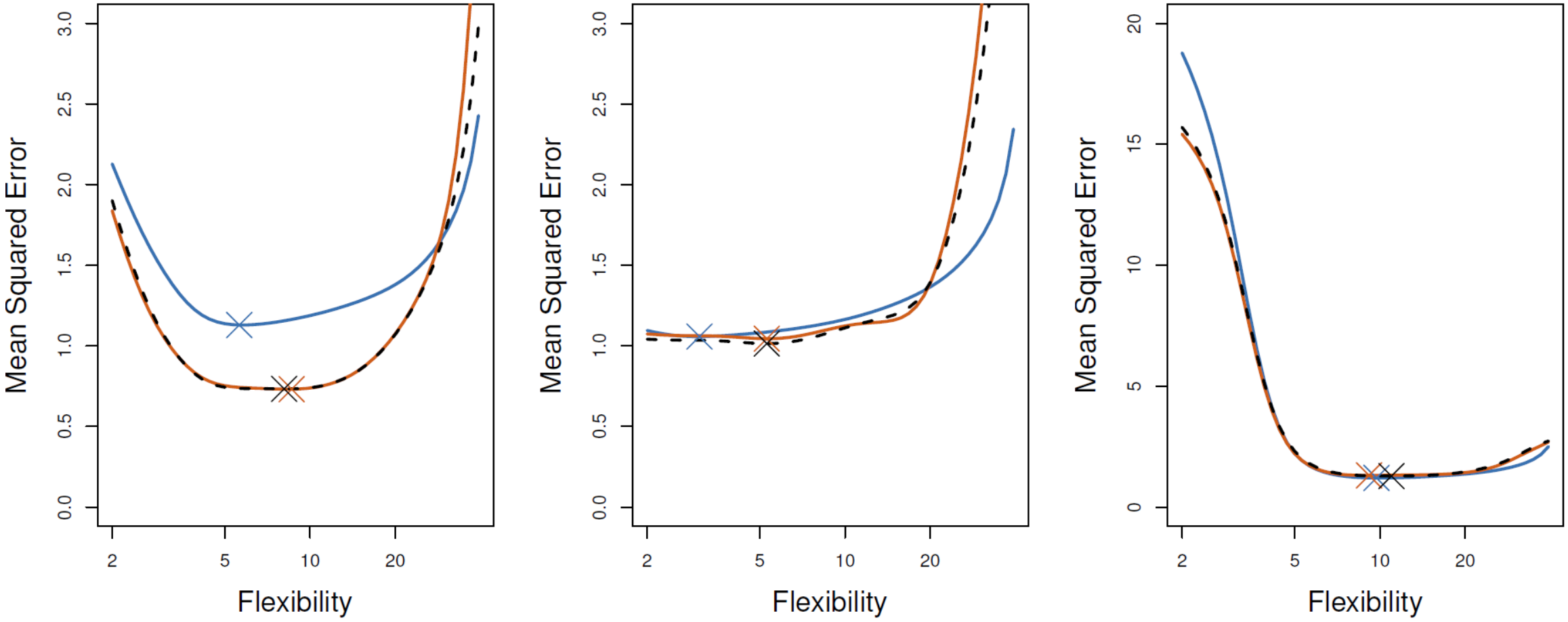




Auto data. Cross-validation was used to estimate the test error that results from predicting mpg using polynomial functions of horsepower

Cross-validation can be used to

- Estimate the test error associated with a given statistical learning method (model assessment)
- Select the appropriate level of flexibility (model selection)



Simulated data sets. The true test MSE is shown in blue, the LOOCV estimate is shown as a black dashed line, and the 10-fold CV estimate is shown in orange

Choice of k

- $k = n$ gives LOOCV
- In practice, $k = 5$ or $k = 10$
  - Computation
  - Bias-variance trade-off

Cross-validation for classification

- Use the number of misclassified observations to quantify test error

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

The Bootstrap

Best investment allocation

- Invest a fraction α of our money in X, and the remaining $1-\alpha$ in Y
- Choose α to minimize the total risk, or variance, of our investment

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

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

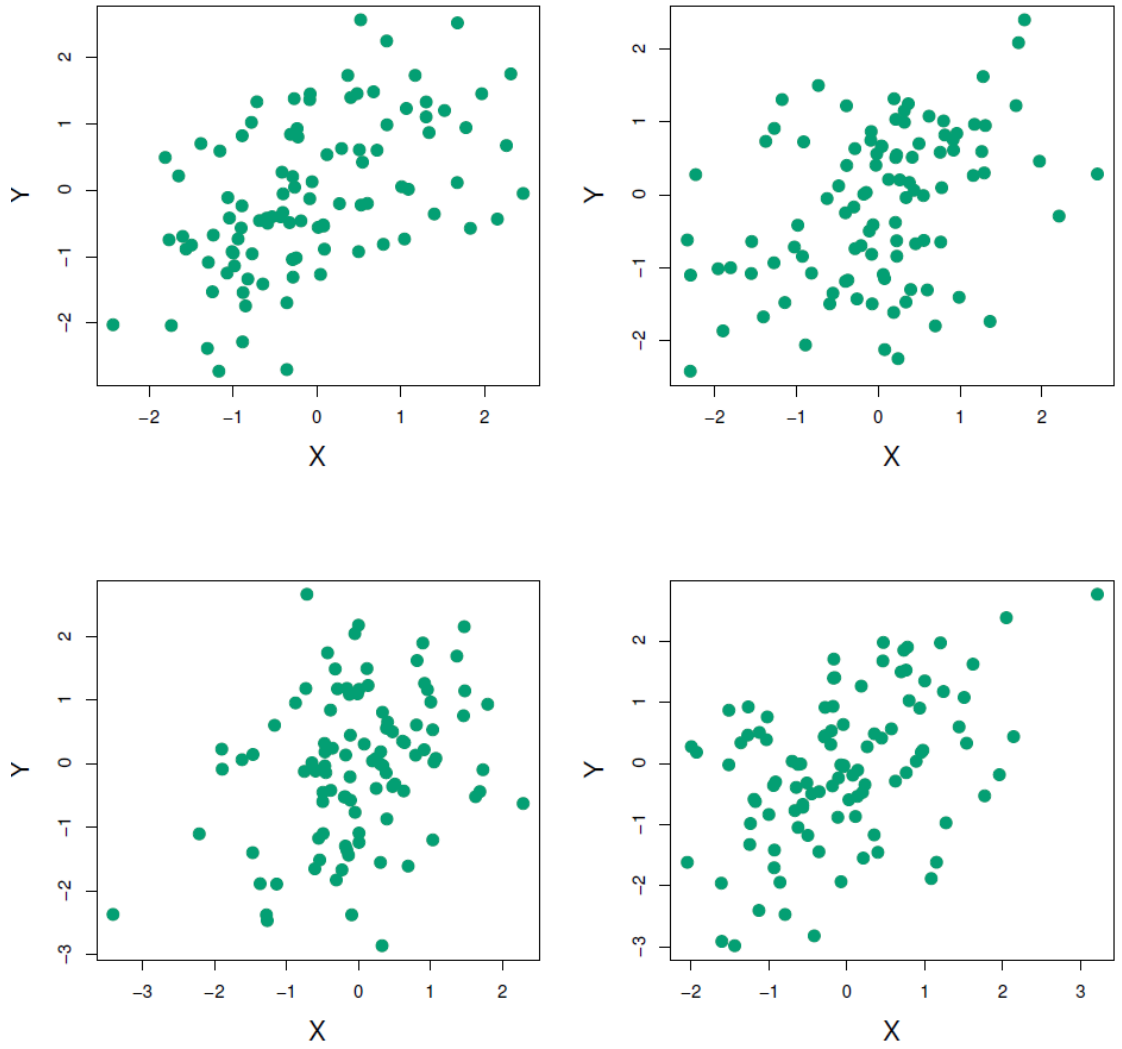
$$\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}}$$

- How can we quantify the accuracy of $\hat{\alpha}$?

The idealized estimate

- Repeat the process of simulating paired observations of X and Y, B times
  - Obtain B estimates for α, $\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_B$
- Calculate

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

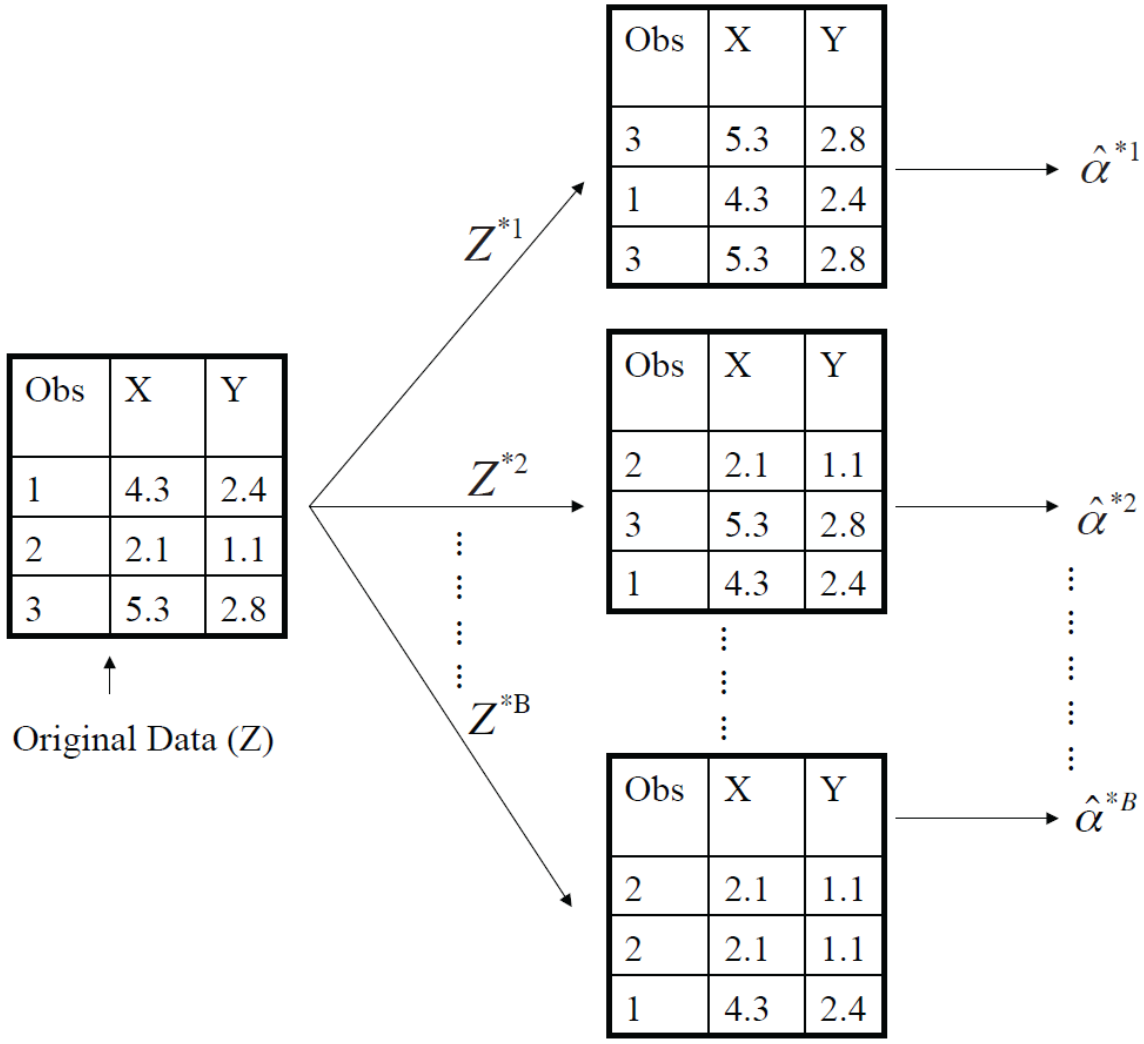


Each panel displays 100 simulated returns for X and Y. The resulting estimates for $\alpha = 0.6$ are 0.576, 0.532, 0.657, and 0.651

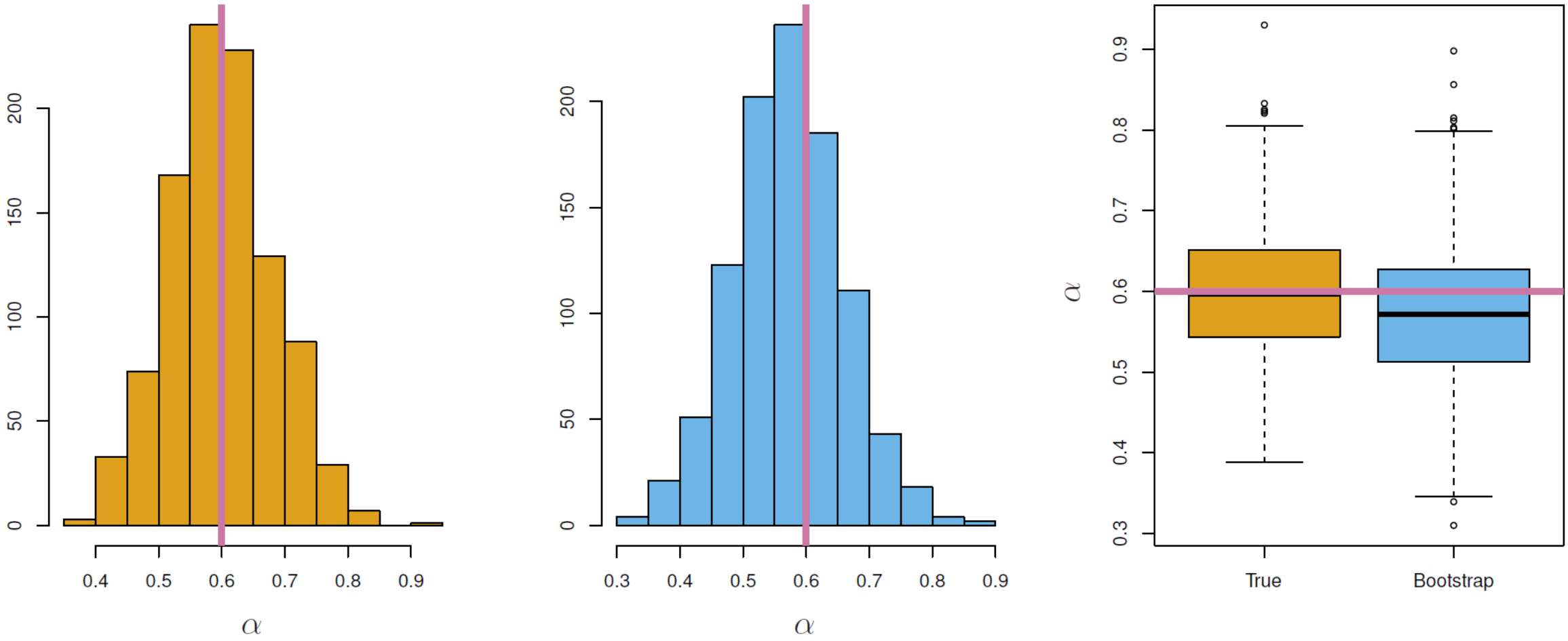
The bootstrap approach

- Repeatedly sample observations from the original data set with replacement, B times
  - Obtain B bootstrap estimates, $\hat{\alpha}_1^*, \hat{\alpha}_2^*, \ldots, \hat{\alpha}_B^*$
- Calculate

$$\bar{\alpha}^* = \frac{1}{B}\sum_{r=1}^{B}\hat{\alpha}_r^*, \mathrm{SE}^*(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}(\hat{\alpha}_r^* - \bar{\alpha}^*)^2}$$



A graphical illustration of the bootstrap approach on a smallsample containing $n = 3$ observations



Histograms of the estimates of α obtained by generating 1,000 simulated data sets from the true population (left) and of the estimates of α obtained from 1,000 bootstrap samples from a single data set (center)