Validation set approach

Goal: Estimate the test error for a supervised learning method.

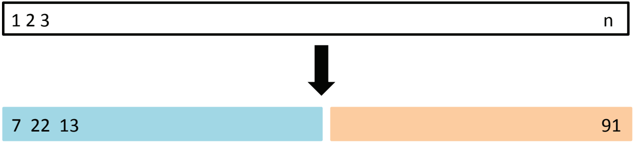
Strategy:

Validation set approach

Goal: Estimate the test error for a supervised learning method.

Strategy:

- Split the data in two parts.

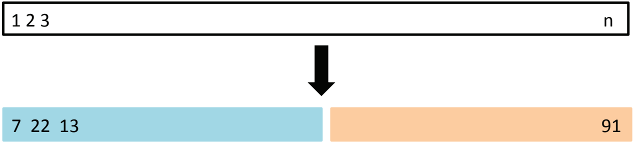


Validation set approach

Goal: Estimate the test error for a supervised learning method.

Strategy:

- Split the data in two parts.
- Train the method in the first part.

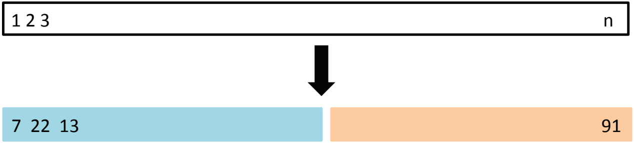


Validation set approach

Goal: Estimate the test error for a supervised learning method.

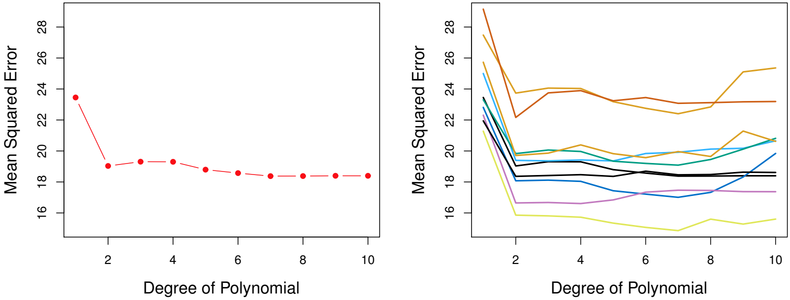
Strategy:

- Split the data in two parts.
- Train the method in the first part.
- Compute the error on the second part.



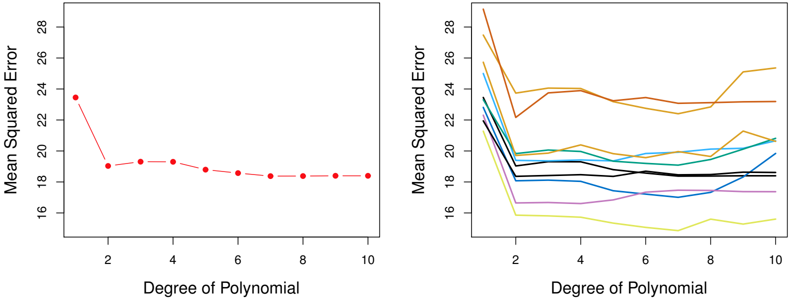
Validation set approach

Polynomial regression to estimate `mpg` from `horsepower` in the Auto data.



Validation set approach

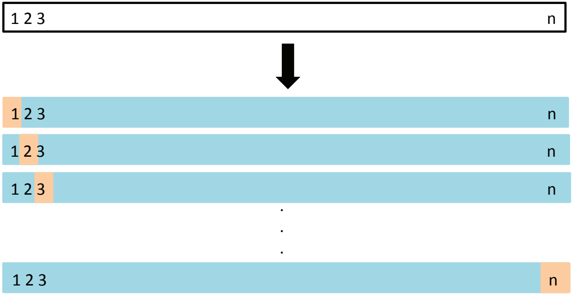
Polynomial regression to estimate `mpg` from `horsepower` in the Auto data.



Problem: Every split yields a different estimate of the error.

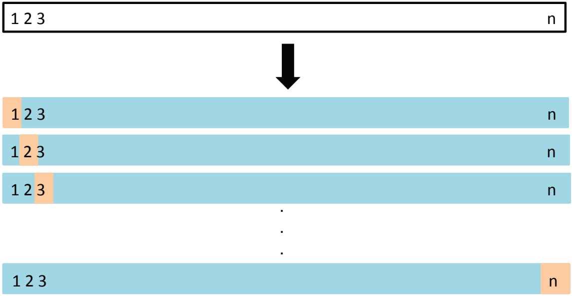
Leave one out cross-validation

- For every $i = 1, \ldots, n$:
  - train the model on every point except i,
  - compute the test error on the held out point.



Leave one out cross-validation

- For every $i = 1, \ldots, n$:
  - train the model on every point except i,
  - compute the test error on the held out point.
- Average the test errors.



Leave one out cross-validation

- For every $i = 1, \ldots, n$:
  - train the model on every point except i,
  - compute the test error on the held out point.
- Average the test errors.

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

Prediction for the i sample without using the ith sample.

Leave one out cross-validation

- For every $i = 1, \ldots, n$:
  - train the model on every point except i,
  - compute the test error on the held out point.
- Average the test errors.

$$\mathsf{CV}_{(n)} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(y_i \neq \hat{y}_i^{(-i)})$$

... for a classification problem.

Leave one out cross-validation

Computing $\mathsf{CV}_{(n)}$ can be computationally expensive, since it involves fitting the model n times.

Leave one out cross-validation

Computing $\mathsf{CV}_{(n)}$ can be computationally expensive, since it involves fitting the model n times.

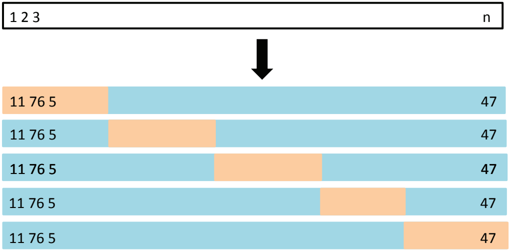
For linear regression, there is a shortcut:

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

where h_{ii} is the leverage statistic.

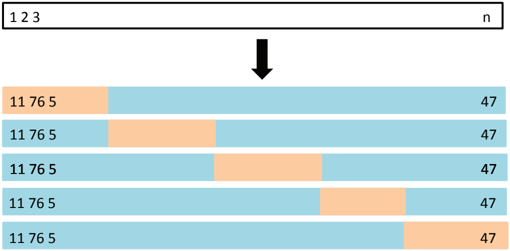
k-fold cross-validation

- Split the data into k subsets or *folds*.



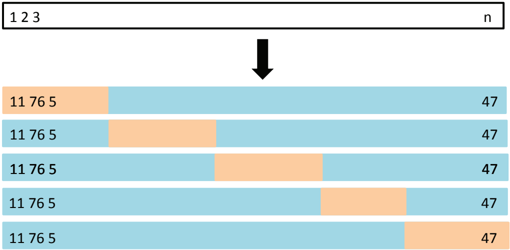
k-fold cross-validation

- Split the data into k subsets or *folds*.
- For every $i = 1, \ldots, k$:
  - train the model on every fold except the ith fold,
  - compute the test error on the ith fold.

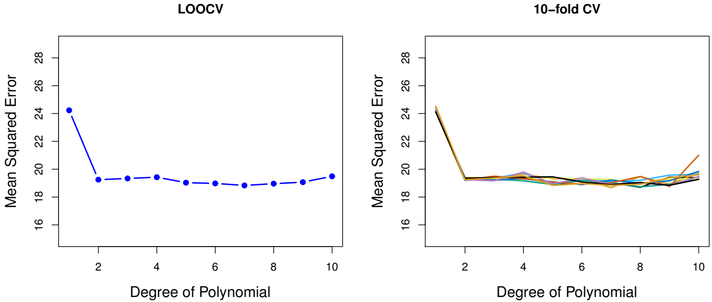


k-fold cross-validation

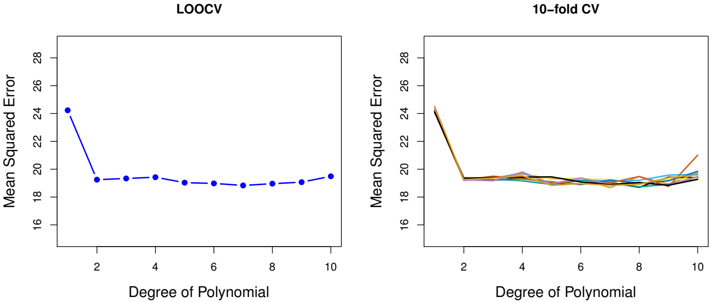
- Split the data into k subsets or *folds*.
- For every $i = 1, \ldots, k$:
  - train the model on every fold except the ith fold,
  - compute the test error on the ith fold.
- Average the test errors.



LOOCV vs. k-fold cross-validation

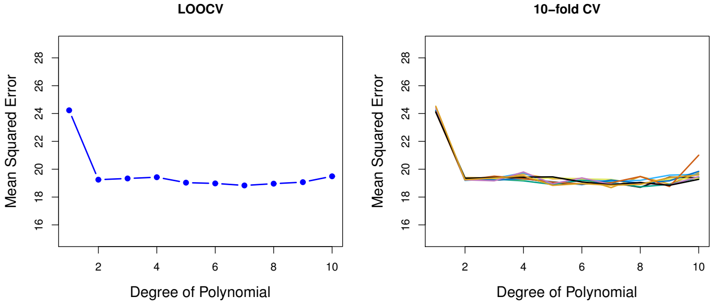


LOOCV vs. k-fold cross-validation



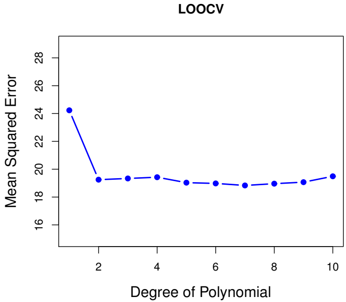
- k-fold CV depends on the chosen split (somewhat).

LOOCV vs. k-fold cross-validation



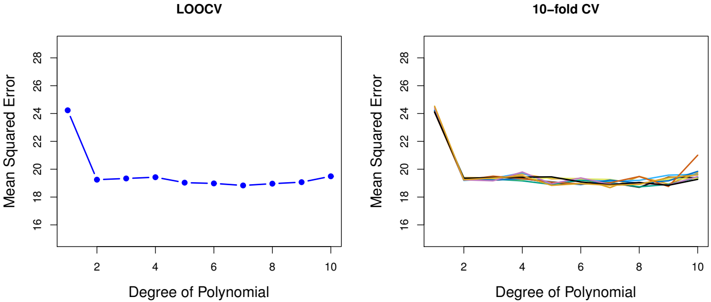
- k-fold CV depends on the chosen split (somewhat).
- In k-fold CV, we train the model on less data than what is available. This introduces **bias** into the estimates of test error.

LOOCV vs. k-fold cross-validation



- k-fold CV depends on the chosen split (somewhat).
- In k-fold CV, we train the model on less data than what is available. This introduces **bias** into the estimates of test error.
- In LOOCV, the training samples highly resemble each other. This increases the **variance** of the test error estimate.

LOOCV vs. k-fold cross-validation



- k-fold CV depends on the chosen split (somewhat).
- In k-fold CV, we train the model on less data than what is available. This introduces **bias** into the estimates of test error.
- In LOOCV, the training samples highly resemble each other. This increases the **variance** of the test error estimate.
- n-fold CV is equivalent LOOCV.