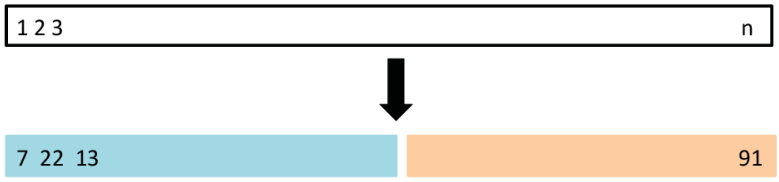

Figure 1: The validation process.

Contents

1 Cross-validation

1.1 Resampling Methods

- **Resampling methods** involve repeatedly drawing samples from a *training* set and refitting a model of interest on each sample in order to obtain additional information about the fitted model.
- **Cross-validation** and the **bootstrap** are the best known.
- They are used for evaluating model's performance (*model assessment*) and selecting the best model (*model selection*).
- For example, they can be used to get the standard deviation and bias of our parameter estimates, as well as estimates of the test error associated with a given statistical learning method, or to select the appropriate level of flexibility.

1.2 Validation-set approach

- Here we randomly divide the available set of samples into two parts: a *training set* and a ***validation or hold-out set***.
- The model is fit on the training set, and 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. This is typically assessed using MSE in the case of a quantitative response and misclassification rate in the case of a qualitative (discrete) response.

The Validation process

A random splitting into two halves: left part is training set, right part is validation set.

Example: the `Auto` data set

`mpg` (gas mileage in miles per gallon) vs `horsepower`.

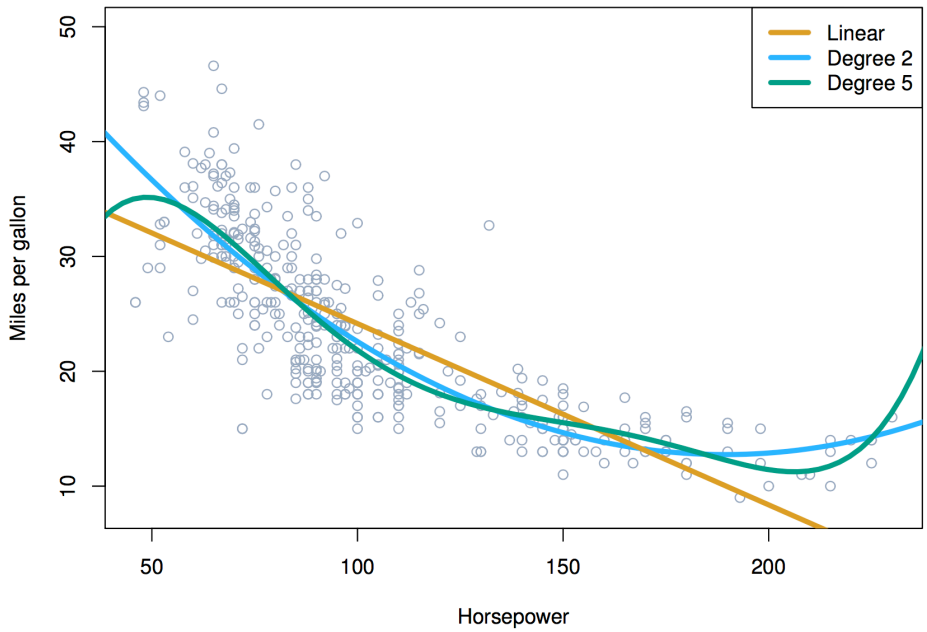

Figure 2: The 'Auto' data set.

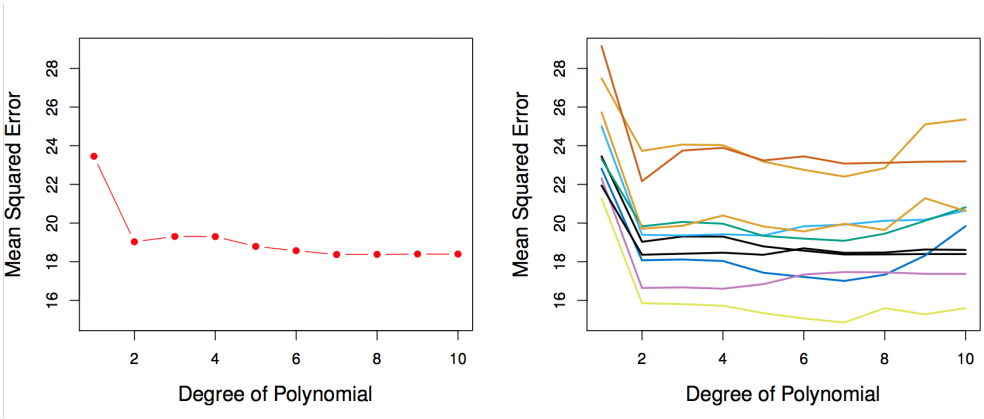

Figure 3: Validation-set approach for the 'Auto' data set.

The data suggest a curved relationship.

We want to compare linear vs higher-order polynomial terms in a linear regression

$$\begin{aligned} \mathrm{M}_2: \ & \mathrm{mpg} = \beta_0 + \beta_1 \times \mathrm{horsepower} + \beta_2 \times \mathrm{horsepower}^2 + \epsilon \\ \mathrm{M}_5: \ & \mathrm{mpg} = \beta_0 + \beta_1 \times \mathrm{horsepower} + \beta_2 \times \mathrm{horsepower}^2 + \ldots \\ & \ldots + \beta_5 \times \mathrm{horsepower}^5 + \epsilon \end{aligned}$$

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.

Left panel shows single split; right panel shows multiple splits.

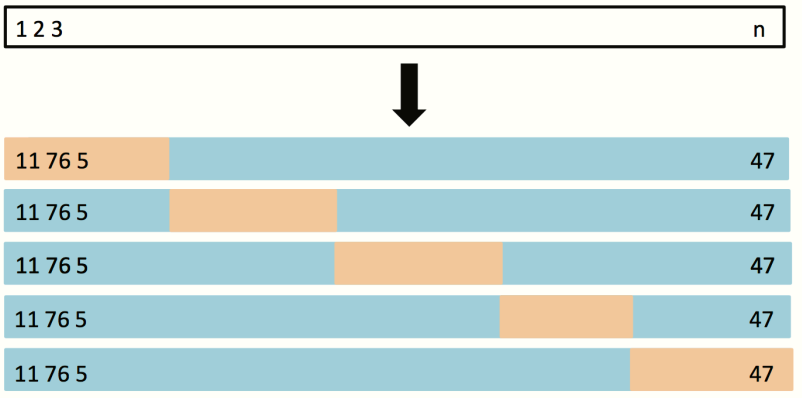

Figure 4: 5-fold Cross-validation.

Drawbacks of validation-set approach

- The validation estimate of the test error can be **highly variable**, depending on precisely which observations are included in the training set and which are in the validation set.
- In the validation approach, only a subset of the observations - those that are included in the training set - are used to fit the model. This suggests that the validation set error may tend to **overestimate** the test error for the model fit on the entire data set. *Why?*

1.3 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 is to randomly divide the data into K equal-sized parts. We leave out part k, fit the model to the other $K-1$ parts (combined), and then 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.

K-fold Cross-validation in detail

Divide the data into K roughly equal-sized parts ($K = 5$ here).

1	2	3	4	5
Validation	Train	Train	Train	Train

- 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

$$\text{CV}_K = \sum_{k=1}^{K} \frac{n_k}{n} \text{MSE}_k$$

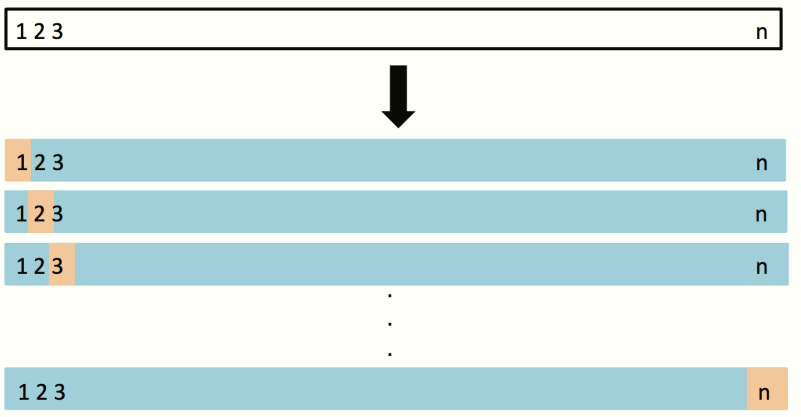

Figure 5: Leave-one-out cross-validation.

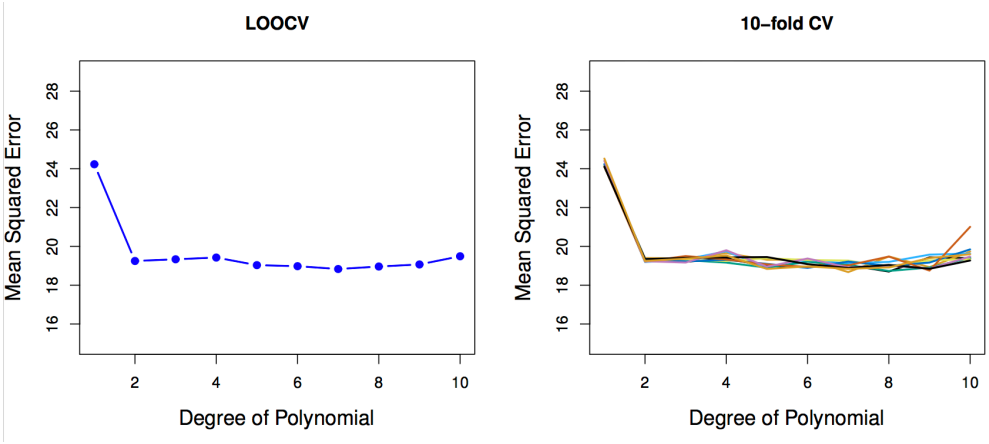

Figure 6: k-fold cv applied on the 'Auto' data set.

where $\text{MSE}_k = \sum_{i \in C_k} (y_i - \hat{y}_i)^2 / n_k$, and $\hat{y}_i$ is the fit for observation i, obtained from the data with part k removed.

- Setting $K = n$ yields n-fold or **leave-one out** cross-validation (LOOCV).

LOOCV

Advantages wrt the validation set approach

- Less bias: LOOCV will give approximately unbiased estimates of the test error
- Performing LOOCV multiple times will always yield the same results

---

10-fold CV were run nine separate times, each with a different random split of the data into ten parts.

Advantages wrt the validation set approach

- Variability is typically much lower
- 5/10-fold CV will lead to an intermediate level of bias

A nice special case!

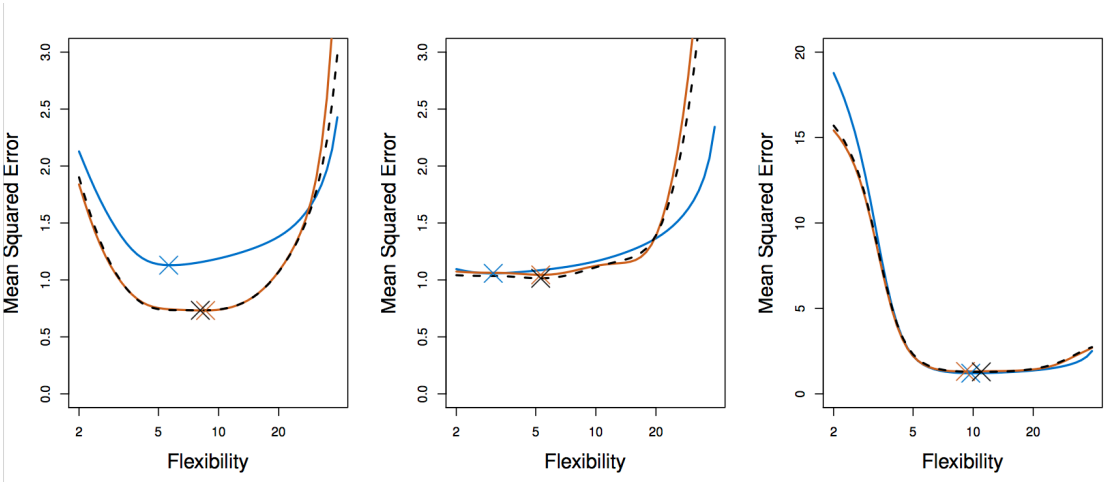

Figure 7: True and estimated test MSE for three simulated data.

- With least-squares linear or polynomial regression, an amazing shortcut makes the cost of LOOCV the same as that of a single model fit! The following formula holds:

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

  where $\hat{y}_i$ is the ith fitted value from the original least squares fit, and h_i is the **leverage** (diagonal of the "hat" matrix). This is like the ordinary MSE, except the ith residual is divided by $1 - h_i$.

LOOCV vs K-fold CV

- LOOCV has the potential to be computationally expensive (except for linear models fit by least squares).
- From the perspective of bias reduction, it is clear that LOOCV is to be preferred to k-fold CV.
- But typically LOOCV doesn't *shake up* the data enough. The estimates from each fold are highly correlated and hence their average can have high variance.
- A better choice is $K = 5$ or 10, as these values have been shown empirically to yield test error estimates that suffer neither from excessively high bias nor from very high variance.

True and estimated test MSE for simulated data

CV estimates and true test error that result from applying smoothing splines to three simulated data sets.

The true test MSE is displayed in blue. The black dashed and orange solid lines respectively show the estimated LOOCV and 10-fold CV estimates

---

Despite the fact that they sometimes underestimate the true test MSE, all of the CV curves come **close to identifying the correct level of flexibility**—that is, the flexibility level corresponding to the smallest test MSE.