14. Model validation

- resampling method
- cross-validation

Resampling methods

- a process of *repeatedly* drawing samples from a training set and refitting a model on each sample
- we seek for information that would not be obtained from fitting the model only *once* using the original training sample
- resampling approaches can be computationally expensive but with nowaday technology, it becomes less prohibitive
  - cross-validation: used in estimation of test error or model flexibility
  - bootstrap: a measure of accuracy of a parameter estimate (not given in this lecture)

Test and Training error rates

- training error rate: the average error that results from using a trained model (or method) back on the training data set
- test error rate: the average error that results from using a statistical learning method to predict the response on a **new observation**
- training error can be quite different from the test error rate
- when test data set is limited, a number of techniques can be used to estimate *test error rate* using the available training data

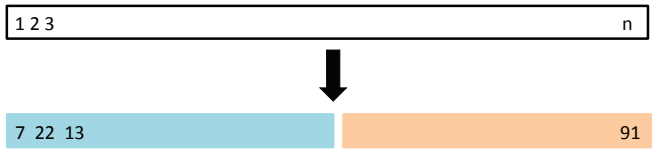
Cross-validation

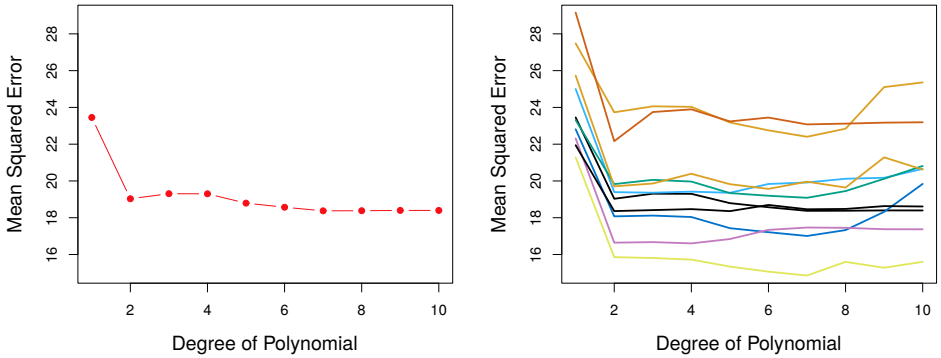
- validation set approach
- leave-one-out cross validation
- k-fold cross validation

Validation set approach

divide available data into two parts:

- training set: used for fitting a model
- validation set: used for predicting the response from the fitted model





- use `Auto` data (fit $y = a_0 + a_1 x + \cdots + a_n x^n$)
- left: validation error from a single split of data
- right: randomly split the training and validation sets; repeat 10 times

- using quadratic term can reduce MSE more considerably than a linear term
- cubic term does not give better prediction than using a quadratic term
- all ten curves on RHS confirm that using higher order than quadratic do not gain a benefit in prediction

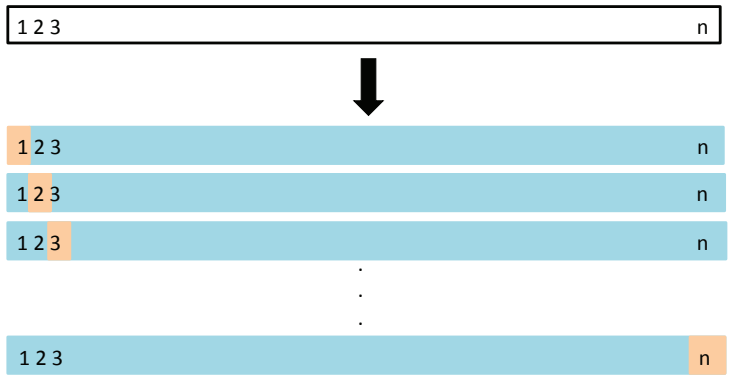
drawbacks of validation set approach

- validation estimate of test error rate can be highly varied, depending on which observations included in the training and validation sets
- it may *overestimate* test error rate for the model fit on the entire data set (because the model is fitted on fewer observations – poorer performance)

Leave-one-out cross validation (LOOCV)

divide available data $\{(x_i, y_i)\}_{i=1}^n$ into two parts:

- training set: $\{(x_2, y_2), \ldots, (x_n, y_n)\}$ (shown in blue)
- validation set: $\{(x_1, y_1)\}$ (shown in beige color)

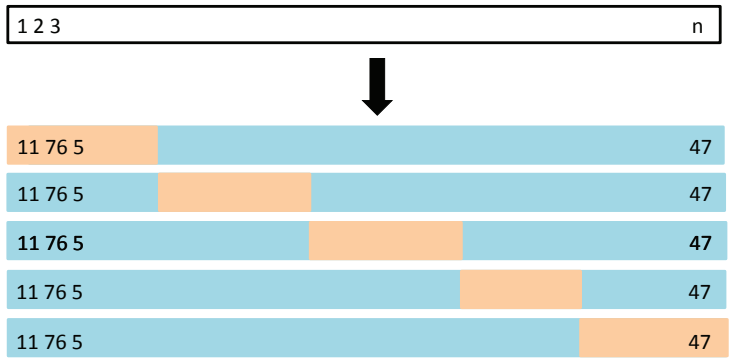


- repeat choosing $\{(x_k, y_k)\}$ as the validation set, where $k = 2, \ldots, n$ and compute MSE_1, $\text{MSE}_2, \ldots, \text{MSE}_n$
- the test error rate is estimated by **averaging** the n MSE's

k-fold cross validation

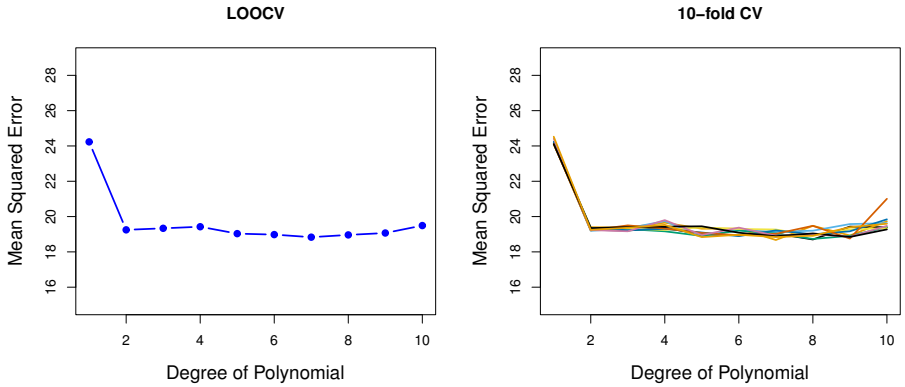
divide available data $\{(x_i, y_i)\}_{i=1}^n$ into k groups or folds:

- validation set: the first fold (shown in beige color)
- training set: the remaining $k-1$ folds (shown in blue)



- repeated k times where each time a different fold is regarded as validation set and compute MSE_1, $\text{MSE}_2, \ldots, \text{MSE}_k$
- the test error rate is estimated by **averaging** the k MSE's

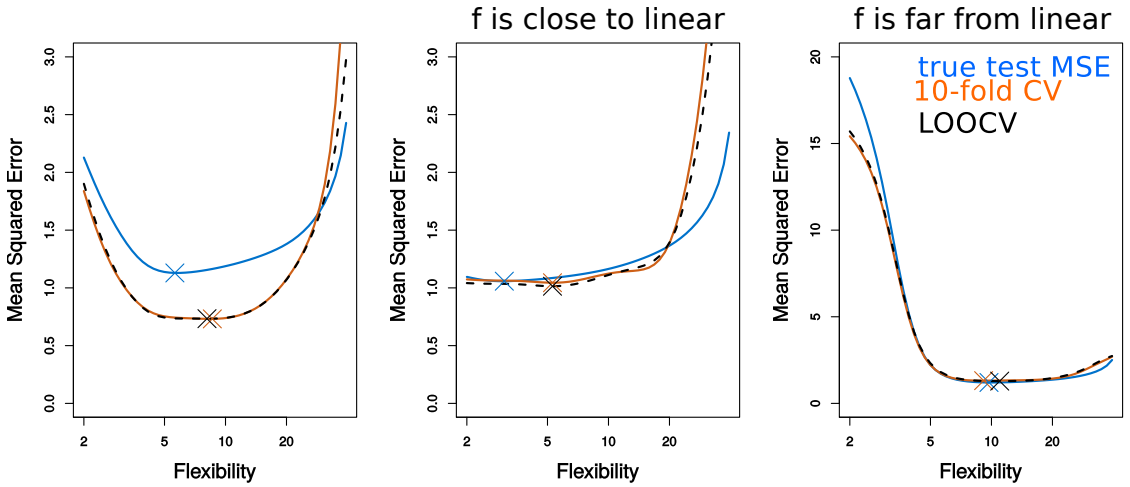
results on `Auto` data set:



properties:

- loocv is a special of k-fold when $k = n$
- loocv requires computation of n times and fitting process can be demanding (if the model is not linear)
- k-fold requires only k times of fitting models; can be more feasible if training process is computationally expensive
- variation of test error from k-fold is less than that of validation set approach

accuracy of test error rate (on simulation data set): using model of smoothing splines



- we can compute the *true test MSE* (assume to know the true description, f) as a function of model complexity
- (left): cv estimates have the correct general U shape but underestimate the test MSE
- (center): cross validation gives oversetimate of test MSE at high flexibility
- (right): the true test MSE and the cv estimates are almost identical

Usage of cross-validation

most of the times we may perform cv on

- a number of statistical methods: and to see which method has the lowest test MSE
- a single statistical method but different flexibilities: and to see which model complexity yield the lowest test MSE

though sometimes cv method underestimate the true test MSE, they can select the correct level of flexibility

Trade-off for k-fold

examine the unbiasedness and variance of test MSE

method	validation set	loocv	k-fold
computation	less	high	feasible
training samples	ratio e.g. 70:30	$n-1$	$(k-1)n/k$
unbiasedness	low	approximately unbiased	intermediate
variance		high	less

- test MSE is calculated by taking the **average** of many MSE's:
- most of MSE's from *loocv* are highly correlated while MSE's of k-fold are less correlated (since loocv uses more overlapped data in training – hence, fitted models are almost identical)
- fact: the sample mean of highly correlated entries has **more variance** than the sample mean of less correlated entries

conclusion: trade-off between bias and variance when choosing k in k-fold

References

All figures and examples are taken from Chapter 5 in

G.James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*, Springer, 2015

Chapter 7 in

T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd edition, Springer, 2009