5 Resampling Methods

- Objection

  1. To detect overfitting
  2. To choose the model with the most suitable flexibility level (model selection)

5.1 Cross-Validation

5.1.1 The Validation Set Approach

- **Implementation**:

  1. Randomly divide observations in half, so $n/2$ as validation set and the other half as training set
  2. Compute the MSE for the validation set
  3. Repeat the steps stated above so we can get the results as Figure 1

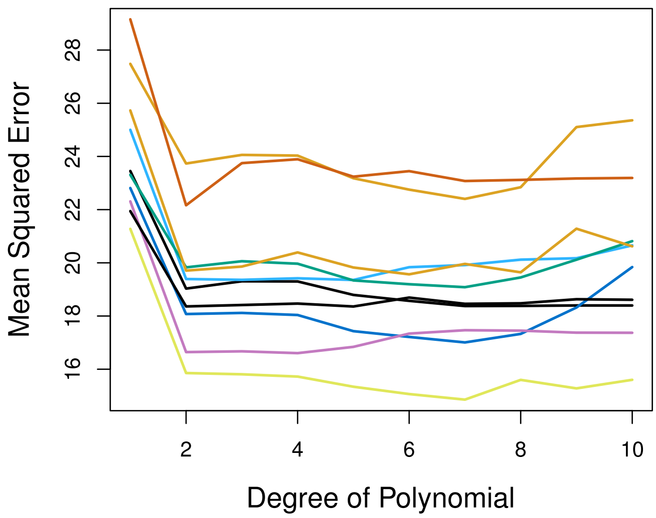

Figure 1

- Advantages

1. Before, we only choose part of the data as training set. Now, we should train our models on different set of data to overcome overfitting.
2. From Figure 1 we observed 10 curves indicate that the model with a quadratic term has a smaller validation set MSE then the model with only a linear term. This helps us to choose the degree of polynomial when training a model.

- Drawbacks

1. The validation estimate of the test error can be highly variable, depending on which observations are included in the training set and in the validation set.
2. In the validation approach only half of the data are used to fit the model. Since statistical methods tend to perform worse when trained on fewer observations. This implies the validation set error rate may tend to overestimate the test error rate for the model fit on the entire data set.

5.2 Leave-One-Out Cross-Validation (LOOCV)

Implementation:

Validation Set	Training Set	MSE
(x_1, y_1)	$\{(x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)\}$	$\text{MSE}_1 = (y_1 - \hat{y}_1)^2$
(x_2, y_2)	$\{(x_1, y_1), (x_3, y_3), \ldots, (x_n, y_n)\}$	$\text{MSE}_2 = (y_2 - \hat{y}_2)^2$
(x_3, y_3)	$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$	$\text{MSE}_3 = (y_3 - \hat{y}_3)^2$
$\vdots$	$\vdots$	$\vdots$
(x_n, y_n)	$\{(x_1, y_1), (x_2, y_2), \ldots, (x_{n-1}, y_{n-1})\}$	$\text{MSE}_n = (y_n - \hat{y}_n)^2$
		$\text{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \text{MSE}_i$

- Advantages

1. LOOCV create less bias. Since we repeatedly fit the method using training sets that contain $n - 1$ observations.

2. LOOCV does not overestimate the test error rate as much as the validation set approach does.

3. LOOCV will always yield the same results compared to validation set approach.

- Drawbacks

1. Computational expensive

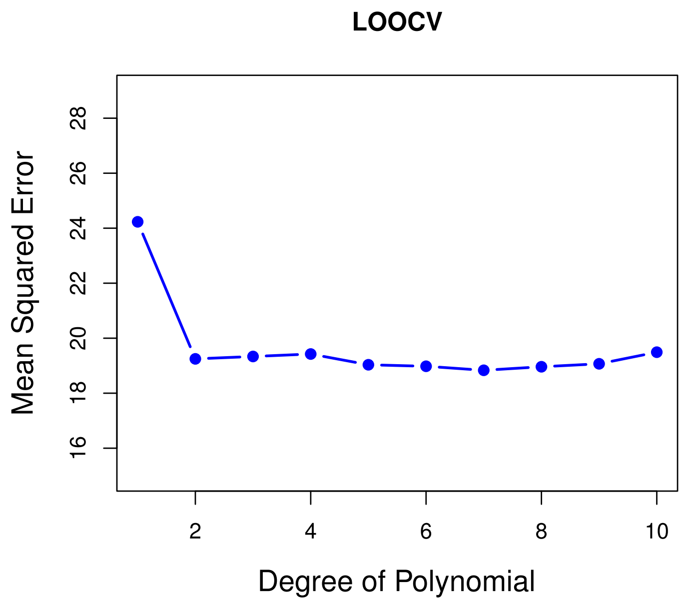

Figure 2: The LOOCV error curve

5.2.1 k-Fold Cross-Validation

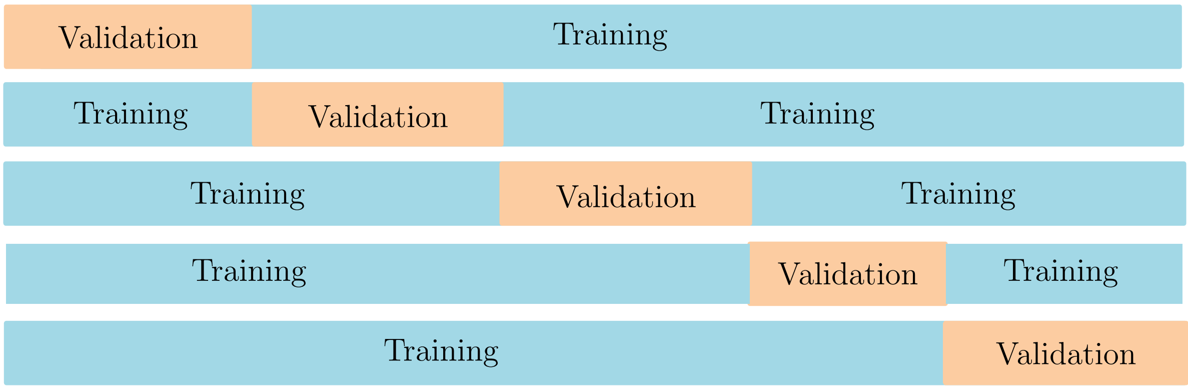

Figure 3: 5-fold CV. A set of n observations is randomly split into five non-overlapping groups. If we set $k = n$, then we get LOOCV.

- **Implementation**:
  1. $k = (1, \ldots, n/k)$
  2. Randomly divide the data into k groups, each group contains n/k observations.
  3. Choose the kth fold as validation set, the other observations as training set.
  4. Compute the MSE for the kth fold.
  5. Repeat the process for k times.

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

When we perform cross-validation, our goal might be to determine how well a method can be expected to perform on independent data; in this case, the actual estimate of the test MSE is of interest. But at other times we are interested in finding the right flexibility level that fit the real data. That means, we are looking for the flexibility level location of the minimum point in the estimated test MSE curve.

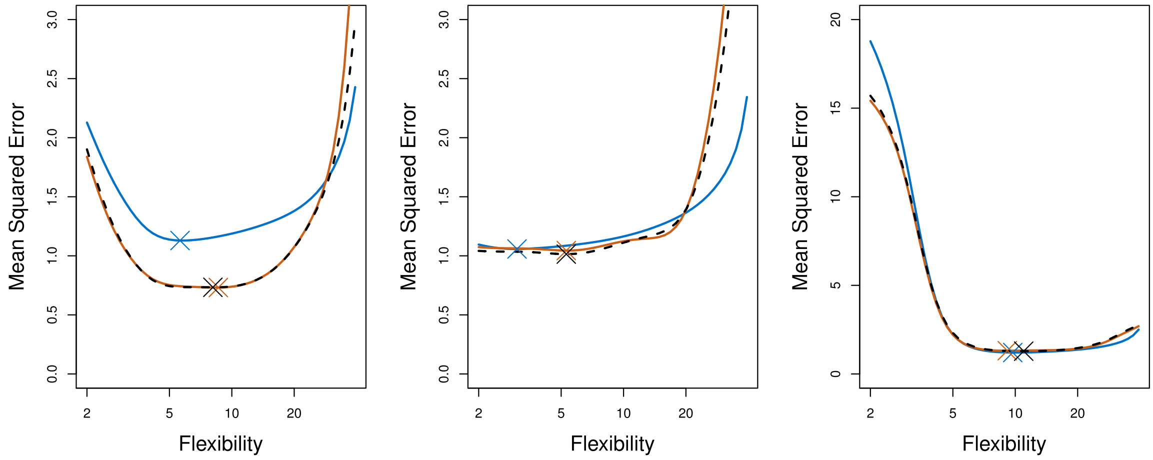

Figure 4: All of the curves come close to identifying the correct levels of flexibility (the flexibility level corresponding to the smallest MSE)

Reference

張翔 (2012), 《提綱契領學統計》, 四版, 鼎茂圖書

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Springer, Berlin: Springer series in statistics.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 6). New York: springer.

Wasserman, L. (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.

J Li, Linear Discriminant Analysis, `http://sites.stat.psu.edu/~jiali/course/stat597e/notes2/lda.pdf`