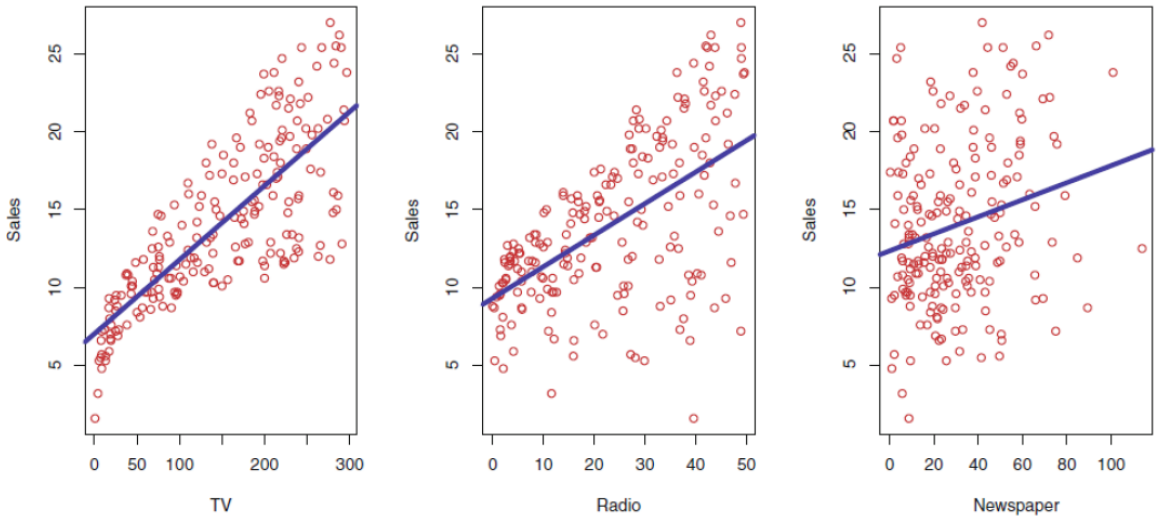
Ⅱ. Statistical Learning

2.1 What is Statistical Learning?

- Simple example
  - ✓ Advertising data set consists of the sales of that product in 200 different markets
  - ✓ Three different media; TV, radio, newspaper
  - ✓ Goal : develop an accurate model that can be used to predict sales on the basis of the three media budgets

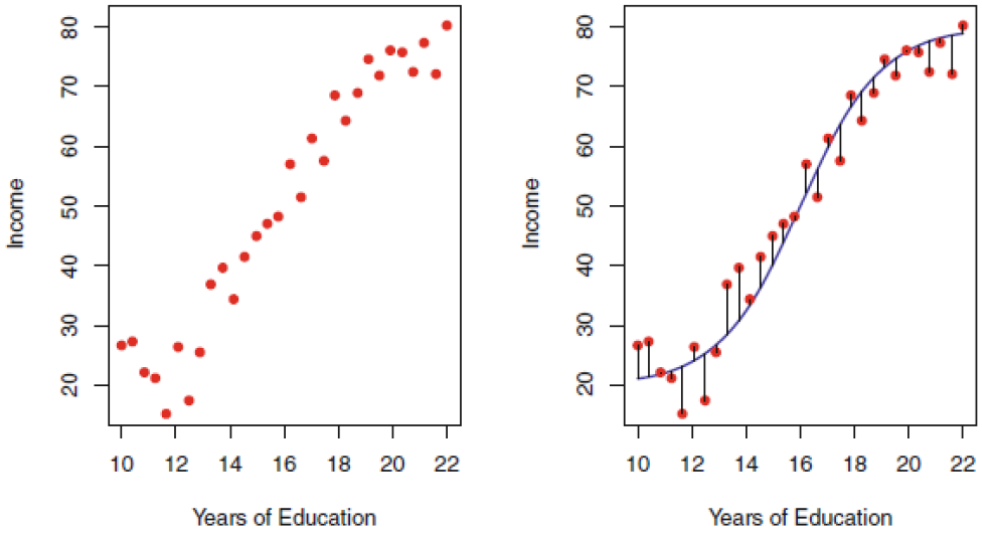


  - ✓ The advertising budgets are input variables. (=predictors, independent variables, features, variables) and X_1 = **TV,** X_2 = **Radio,** X_3 = **Newspaper**
  - ✓ The sales are an output variable. (=response, dependent variables)
- We observe a quantitative response Y and p different predictors, $X_1, X_2, \dots, X_p$
- We assume that there is some relationship between Y and $X = (X_1, X_2, \dots, X_p)$ which can be written in the very general form

$$Y = f(X) + \epsilon$$

- Here f is some fixed but unknown function of $X_1, X_2, \dots, X_p$
- ε is a random **error term**, which is independent of X and has mean zero
- In this formulation, f represents the systematic information that X provides about Y.

- **Another example – income as a function of years of education**



  - One must estimate f based on the observed points.
  - The blue curve represents f and the vertical lines represent the error terms ϵ.
  - In essence, **statistical learning** refers to a set of approaches for estimating f.

2.1.1 Why Estimation f?

- **Prediction**

  - ✓ Since the error term averages to zero, we can predict Y using

$$\hat{Y} = \hat{f}(X)$$

  - ✓ Where $\hat{f}$ represents our estimate for f, and $\hat{Y}$ represents the resulting prediction for Y.

  - ✓ $\hat{f}$ is often treated as a black box, in the sense that one is not typically concerned with the exact form of f, provided that it yields accurate predictions for Y.

  - ✓ Accuracy of $\hat{Y}$ as a prediction for Y depends on two quantities

    1) Reducible Error

    Estimation 과정에서 적절한 통계적 테크닉으로 줄일 수 있는 에러

    2) Irreducible Error

    X를 통해서 Y를 예측할 수 없는 에러. ϵ 의 존재 이유.

  - ✓ Assume for a moment that both $\hat{f}$ and X are fixed. Then, it is easy to show that

$$\begin{aligned} E(Y - \hat{Y})^2 &= E[f(X) + \epsilon - \hat{f}(X)]^2 \\ &= [f(X) - \hat{f}(X)]^2 + Var(\epsilon) \end{aligned}$$

  - ✓ Where $\mathbf{E(Y - \hat{Y})^2}$ represents the average, or **expected value**, of the squared difference between the predicted and actual value of Y

  - ✓ $[f(X) - \hat{f}(X)]^2$: **Reducible Error**

  - ✓ $\mathrm{Var}(\epsilon)$: **Irreducible Error**

The focus of this book is on techniques for estimating with the aim of **minimizing the reducible error.** Irreducible error will always provide an upper bound on the accuracy of our prediction for Y. This bound is almost always unknown in practice.

- **Inference**
  - ✓ Understand the relationship between X and Y, or more specifically, to understand how Y changes as a function of $X_1, X_2, \ldots, X_p$
  - ✓ Now $\hat{f}$ cannot be treated as a black box, because we need to know its exact form.
  - ✓ Questions

    Which predictors are associated with the response?

    What is the relationship between the response and each predictor?

    Can the relationship between and each predictor be adequately summarized using a linear equation, or is the relationship more complicated?

Depending on whether our ultimate goal is prediction, inference, or a combination of the two, different methods for estimating f may be appropriate.

2.1.2 How Do We estimate f?

1. **Training data** : Observations we use to train, or teach, our method how to estimate
2. Let x_{ij} represent the value of the jth predictor, or input, for observation i, where i = 1,2, ..., n and j = 1, 2, ..., p
3. Correspondingly, let y_i represent response variable for the i th observation
4. Then our training data consist of $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ where $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$
5. **Goal :** find a function $\hat{f}$ such that Y ≒ $\hat{f}(X)$ for any observation (X, Y)

Parametric Methods

- make an assumption about the functional form, or shape, of $\boldsymbol{f}$

  Example : linear model

$$\mathrm{f}(\mathbf{X}) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 X_1} + \boldsymbol{\beta_2 X_2} + \cdots + \boldsymbol{\beta_p X_p}$$

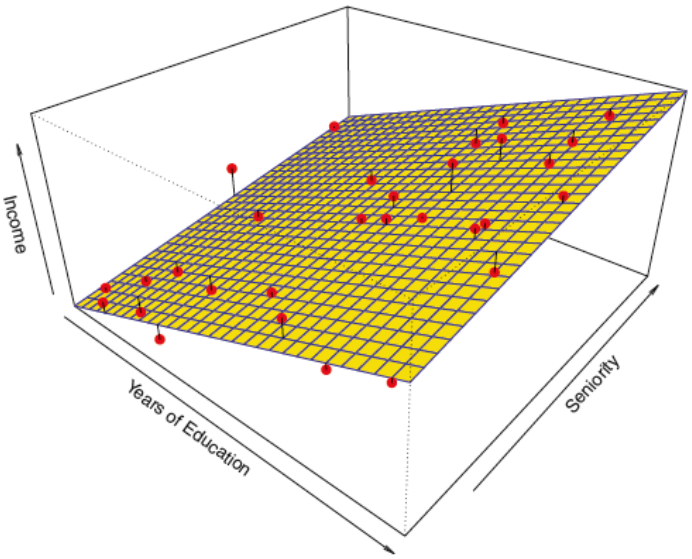
- use the training data to fit or train the model & estimate the parameters

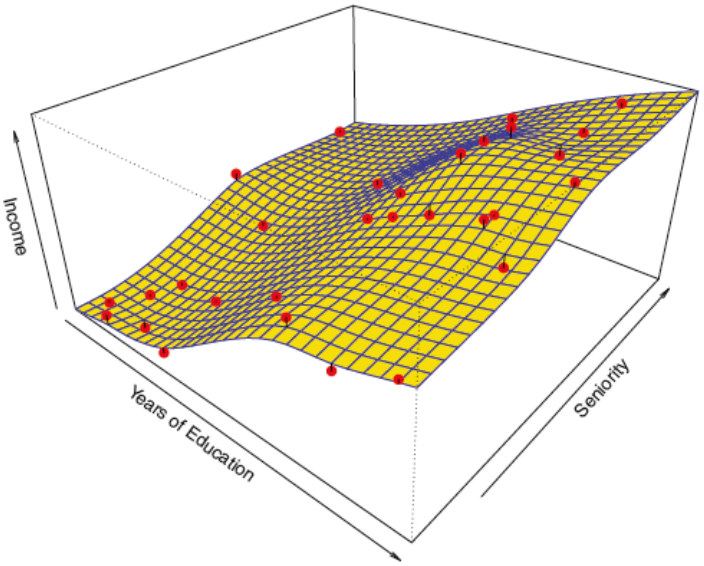
  Example : (ordinary) least squares to estimate $\beta_0, \beta_1, \dots, \beta_p$

- Parametric methods reduce the problem of estimating $\boldsymbol{f}$ down to one of estimating a set of parameters
- Disadvantage: Model we choose will usually not match the true unknown form of $\boldsymbol{f}$

  => can address this problem by choosing flexible models that can fit many different possible functional forms flexible for $\boldsymbol{f}$
- However, fitting a more flexible model requires estimating a greater number of parameters.

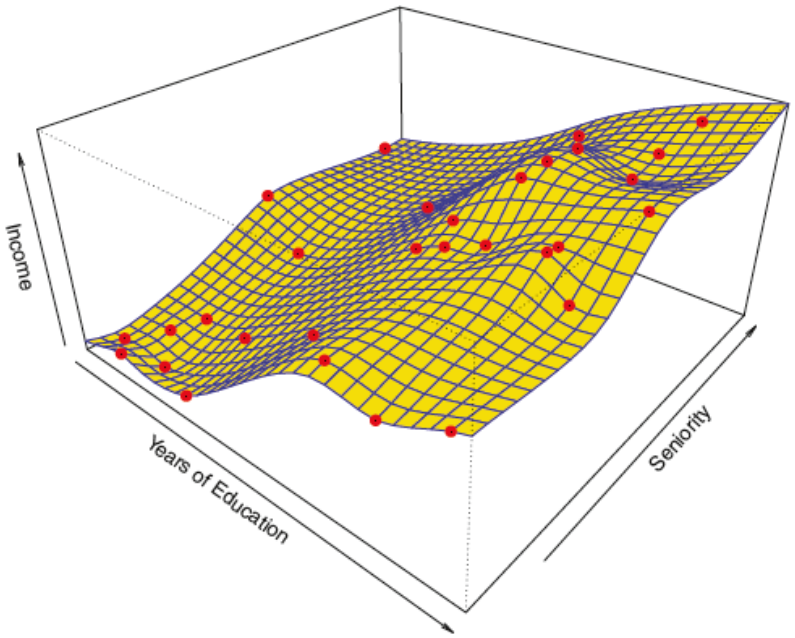
  => **Overfitting problem** : A model follows the errors, or noise, too closely


Income
Years of Education
Seniority


Income
Years of Education
Seniority

Non-Parametric Methods

- do not make explicit assumptions about the functional form of f
- seek an estimate of f that gets as close to the data points as possible without being too rough or wiggly.
- Advantage : potential to accurately fit a wider range of possible shapes for f
- Disadvantage : a very large number of observations is required in order to obtain an accurate estimate for f



2.1.3 The Trade-Off Between Prediction Accuracy and Model Interpretability

- If we are mainly interested in inference, then restrictive models are much more interpretable.
- Very flexible approaches can lead to such complicated estimates of f that it is difficult to understand how any individual predictor is associated with the response.
- Flexible methods에서는 overfitting의 위험성이 있으므로 오히려 less flexible method를 사용하는 것이 더 정확한 predictions을 내릴 수도 있음.

2.1.4 Supervised Versus Unsupervised Learning

- **Supervised learning** : For each observation of the predictor measurement(s) $x_{i,}$ i : 1,2,...,n, there is an associated response measurement y_i

  Many classical statistical learning methods such as linear regression and logistic regression, as well as more modern approaches such as GAM, boosting, and support vector machines, operate in the supervised learning domain.

- **Unsupervised learning** : For every observation $x_{i,}$ i : 1,2,...,n, we observe a vector of measurements x_i but no associated response y_i

  We can seek to understand the relationships between the variables or between the observations. cluster analysis, or clustering

  Visual inspection is simply not a viable way to identify clusters. For this reason, automated clustering methods are important.

- **Semi-supervised learning problem** : observation 중 일부는 predictor와 response가 둘 다 있고, 또 다른 일부는 predictor만 있는 경우

2.1.5 Regression Versus Classification Problems

- Variables can be characterized as either **quantitative** or **qualitative** (**categorical**).
- Problems with a quantitative response → **Regression problems**
- Problems with a qualitative response → **Classification problems**

2.2 Assessing Model Accuracy

2.2.1 Measuring the Quality of Fit

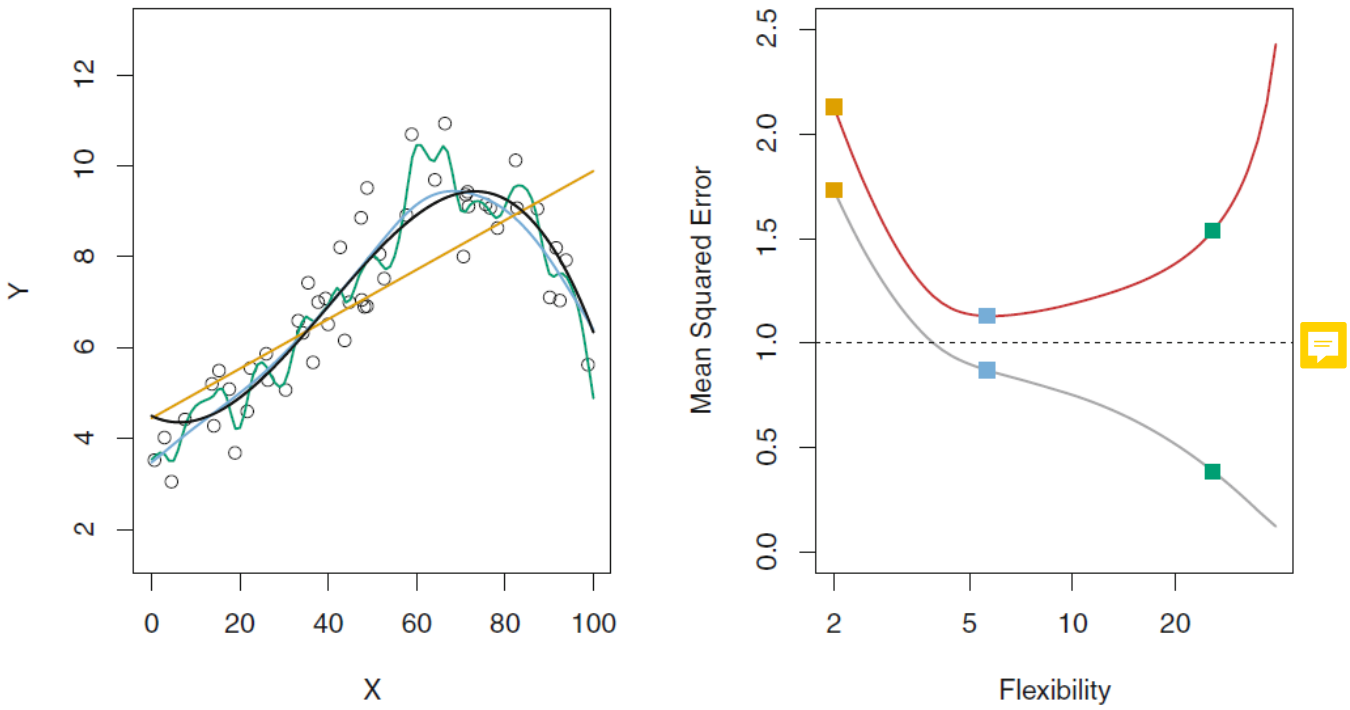
- How well its predictions actually match the observed data
- The most commonly-used measure is the **mean squared error (MSE)**.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

- **Training MSE**: The MSE computed using the training data that was used to fit the model
- 그렇지만 우리가 정말 알고 싶은 것은 **test MSE**. Test MSE가 가장 작은 모델을 선택해야 함.
  - We want to know whether $\hat{f}(x_0)$ is approximately equal to y_0, where (x_0, y_0) is a previously unseen test observation not used to train the statistical learning method.

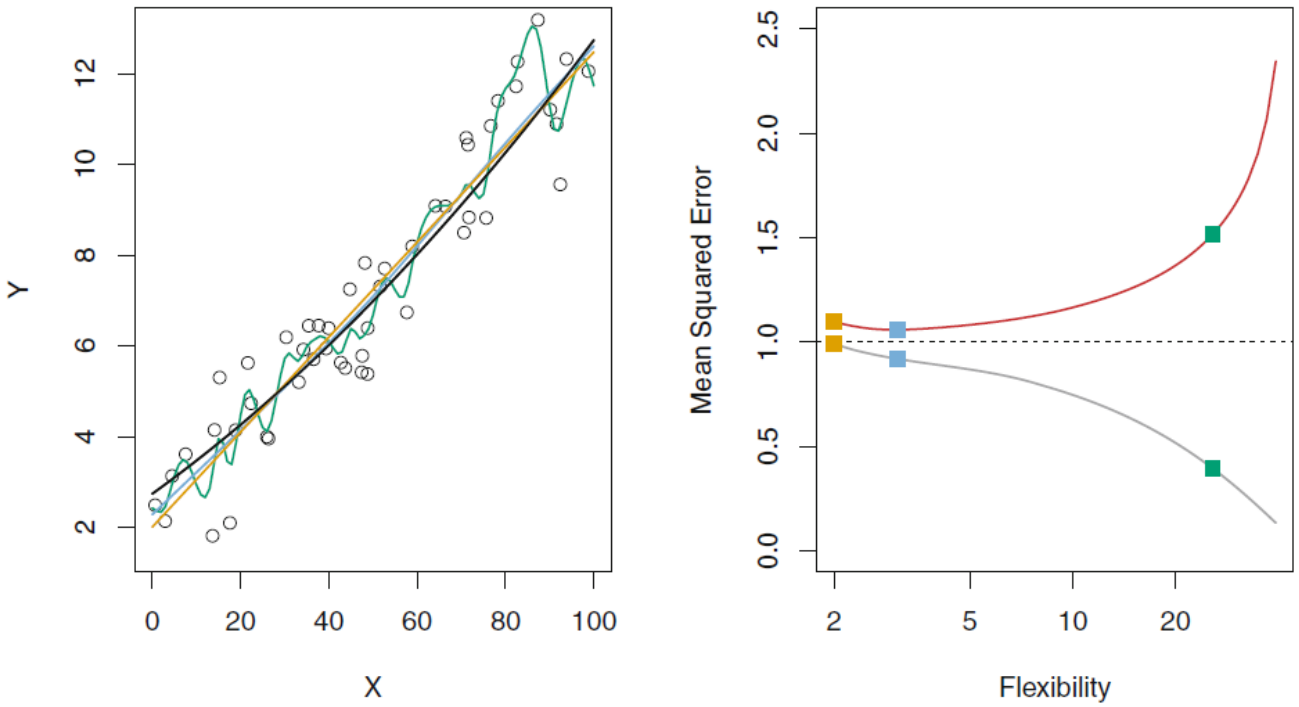
$$Ave(y_0 - \hat{f}(x_0))^2$$

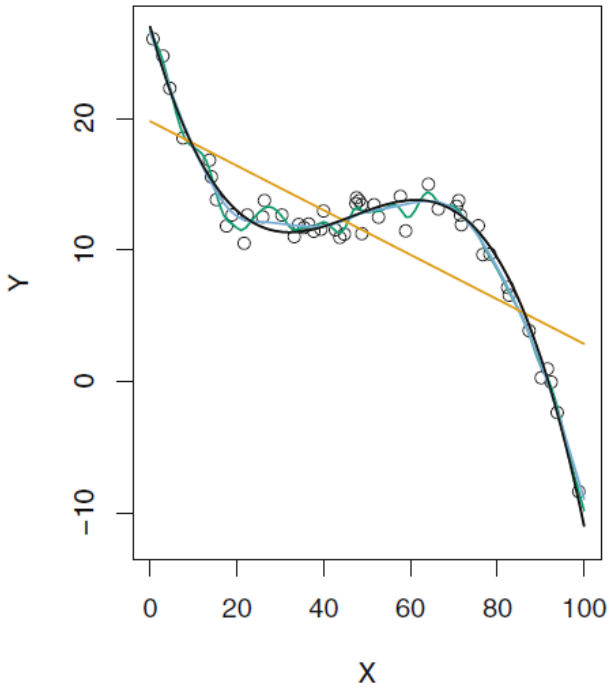
- No guarantee that the method with the lowest training MSE will also have the lowest test MSE.

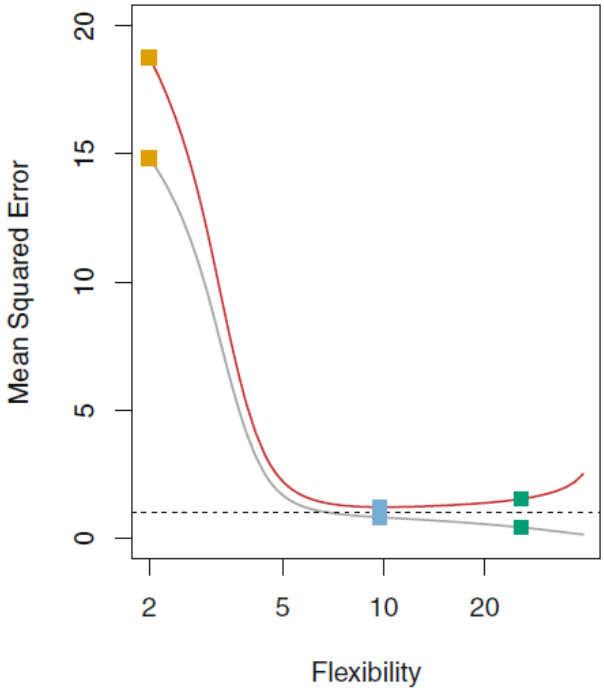


- 왼쪽 그림

    - The black curve: true f
    - The orange, blue and green curves: three possible estimates for f obtained using methods with increasing levels of flexibility. As the level of flexibility increases, the curves fit the observed spline data more closely.
- 오른쪽 그림
    - The grey curve: the average training MSE as a function of flexibility (**degrees of freedom**)
        - The training MSE declines monotonically as flexibility increases.
    - The red curve: the test MSE
        - The test MSE initially declines as the level of flexibility increases. However, at some point the test MSE levels off and then starts to increase again. (U-shaped)
        - The blue curve minimizes the test MSE.
    - The horizontal dashed line: $Var(\epsilon)$, the irreducible error, which corresponds to the lowest achievable test MSE among all possible methods.
    - As the flexibility of the statistical learning method increases, we observe a monotone decrease in the training MSE and a U-shape in the test MSE.
- When a given method yields a small training MSE but a large test MSE, we are said to be **overfitting** the data.
- We almost always expect the training MSE to be smaller than the test MSE.
- Usually no test data are available. $\rightarrow$ **cross-validation**, which is a method for estimating test MSE using the training data





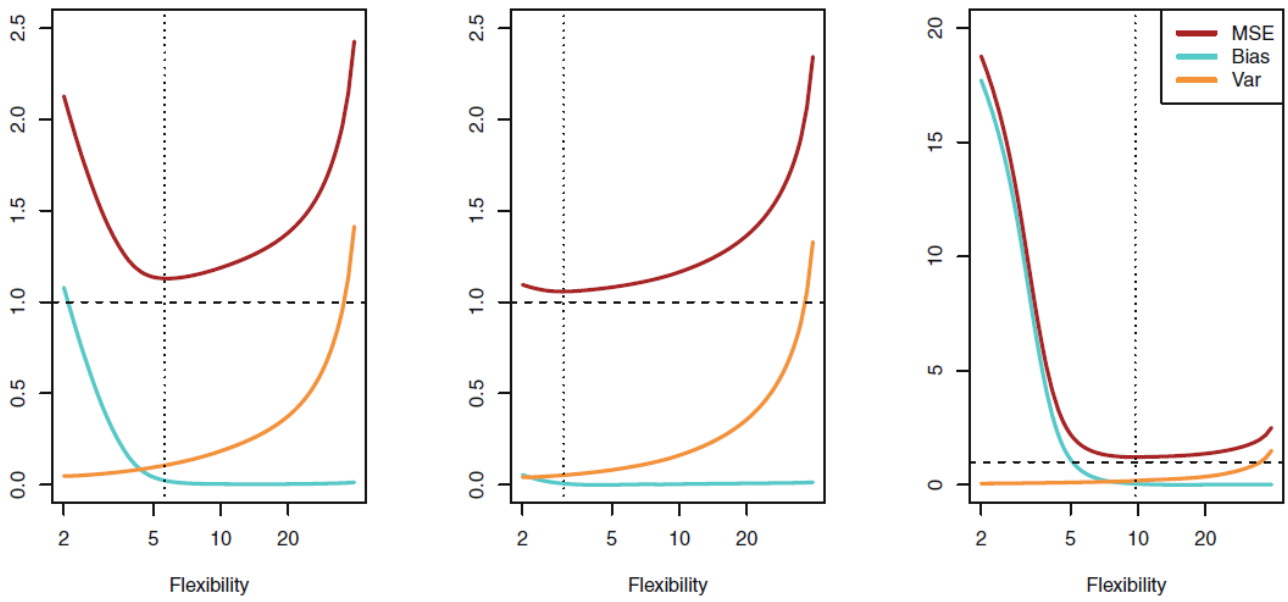


2.2.2 The Bias-Variance Trade-Off

- The expected test MSE, for a given value x_0, can always be decomposed into the sum of three fundamental quantities: the **variance** of $\hat{f}(x_0)$, the squared **bias** of $\hat{f}(x_0)$ and the variance of the error terms ϵ.

$$E(y_0 - \hat{f}(x_0))^2 = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(\epsilon)$$

- $E(y_0 - \hat{f}(x_0))^2$ refers the expected test MSE, the average test MSE that we would obtain if we repeatedly estimated f using a large number of training sets, and tested each at x_0.
- In order to minimize the expected test error, we need to select a statistical learning method that simultaneously achieves **low variance** and **low bias**.
- **Variance** refers to the amount by which $\hat{f}$ would change if we estimated it using a different training data set. Ideally, the estimate for f should not vary too much between training sets.
- **Bias** refers to the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model.
- As a general rule, as we use more flexible methods, the variance will increase and the bias will decrease.



- The horizontal dashed line represents $Var(\epsilon)$, the irreducible error.

2.2.3 The Classification Setting

- Estimate f on the basis of training observations $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, where $y_1, \ldots, y_n$ are qualitative.
- The most common approach for quantifying the accuracy of our estimate $\hat{f}$ is the **training error rate**, the proportion of mistakes that are made if we apply our estimate $\hat{f}$ to the training observations.

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

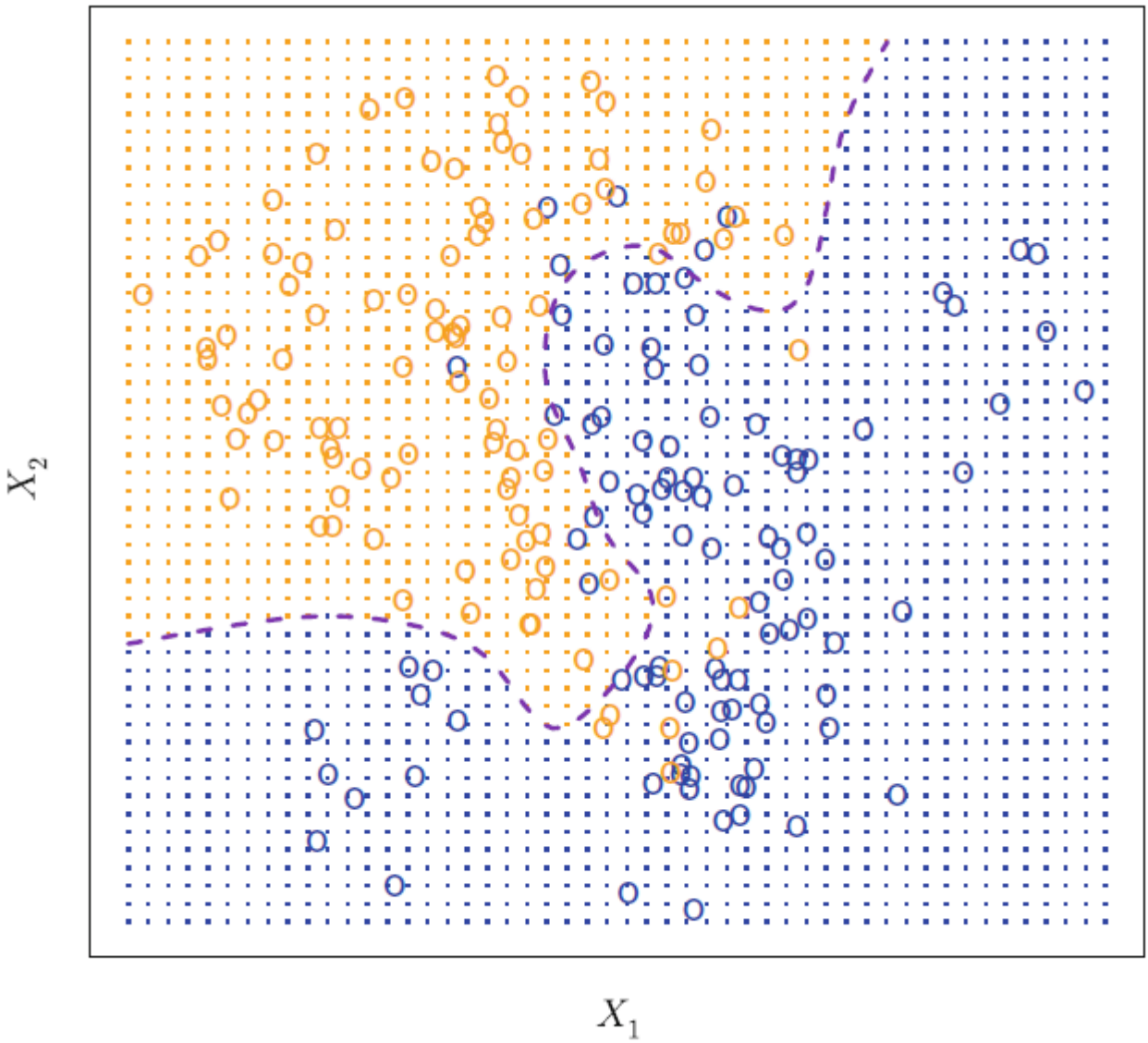
- Here $\hat{y}_i$ is the predicted class label for the ith observation using $\hat{f}$. And $I(y_i \neq \hat{y}_i)$ is an indicator variable that equals 1 if $y_i \neq \hat{y}_i$ and zero if $y_i = \hat{y}_i$.
- The **test error rate** associated with a set of test observations of the form (x_0, y_0) is given by

$$Ave(I(y_i \neq \hat{y}_i))$$

- where $\hat{y}_0$ is the predicted class label that results from applying the classifier to the test observation with predictor x_0.

The Bayes Classifier

- The test error rate is minimized, on average, by a very simple classifier that assigns each observation to the most likely class, given its predictor values.
- We should simply assign a test observation with predictor vector x_0 to the class j for which $Pr(Y = j|X = x_0)$ is largest.
- In a two-class problem where there are only two possible response values, say class 1 or class 2, the Bayes classifier corresponds to predicting class one if $Pr(Y = 1|X = x0) > 0.5$, and class two otherwise.



- A simulated data set in a two-dimensional space consisting of predictors X_1 and X_2.
  - The orange and blue circles correspond to training observations that belong to two different classes.
  - The orange shaded region reflects the set of points for which $Pr(Y = orange|X)$ is greater than 50%, while the blue shaded region indicates the set of points for which the probability is below 50%.
  - The purple dashed line represents the points where the probability is exactly 50%. $\rightarrow$ **Bayes decision boundary**
- The Bayes classifier produces the lowest possible test error rate. $\rightarrow$ **Bayes error rate**
  - The error rate at $X = x_0$ will be $1 - max_j Pr(Y = j|X = x_0)$.
  - In general, the overall Bayes error rate is given by $1 - E(max_j Pr(Y = j|X))$.
- The Bayes error rate is analogous to the irreducible error.

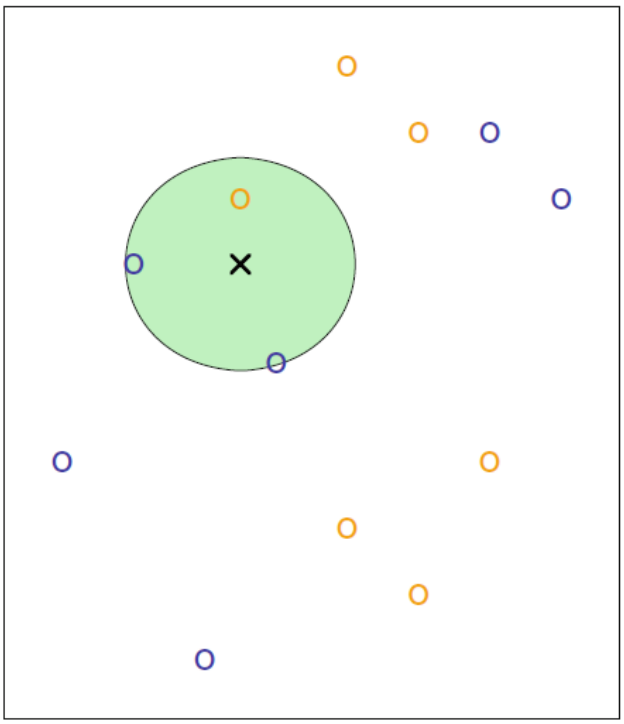
K-Nearest Neighbors

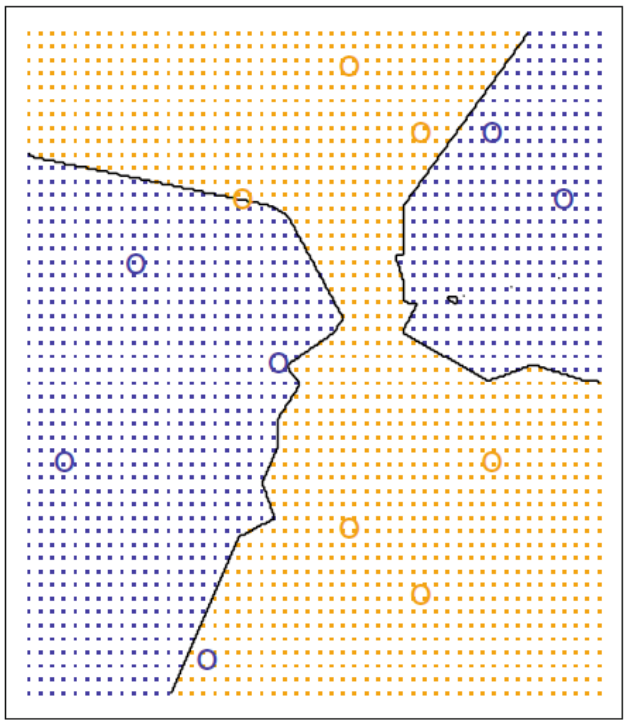
- For real data, we do not know the conditional distribution of Y given X, and so computing the Bayes classifier is impossible.
- The Bayes classifier serves as an unattainable gold standard against which to compare other methods.

- Many approaches attempt to estimate the conditional distribution of Y given X, and then classify a given observation to the class with highest estimated probability.
- K-nearest neighbors (KNN) classifier: identifies the K points in the training data that are closest to x_0, represented by $\mathcal{N}_0$. It then estimates the conditional probability for class j as the fraction of points in $\mathcal{N}_0$ whose response values equal j:

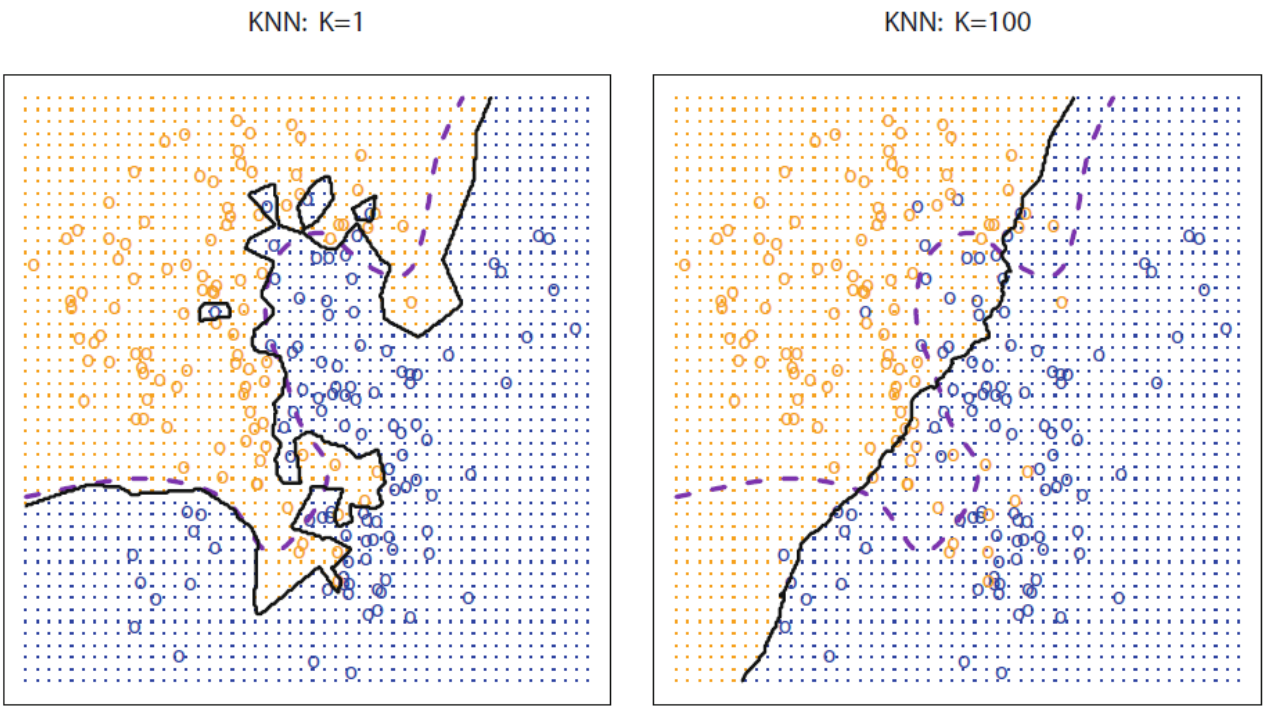
$$Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j)$$

- Finally, KNN applies Bayes rule and classifies the test observation x_0 to the class with the largest probability.

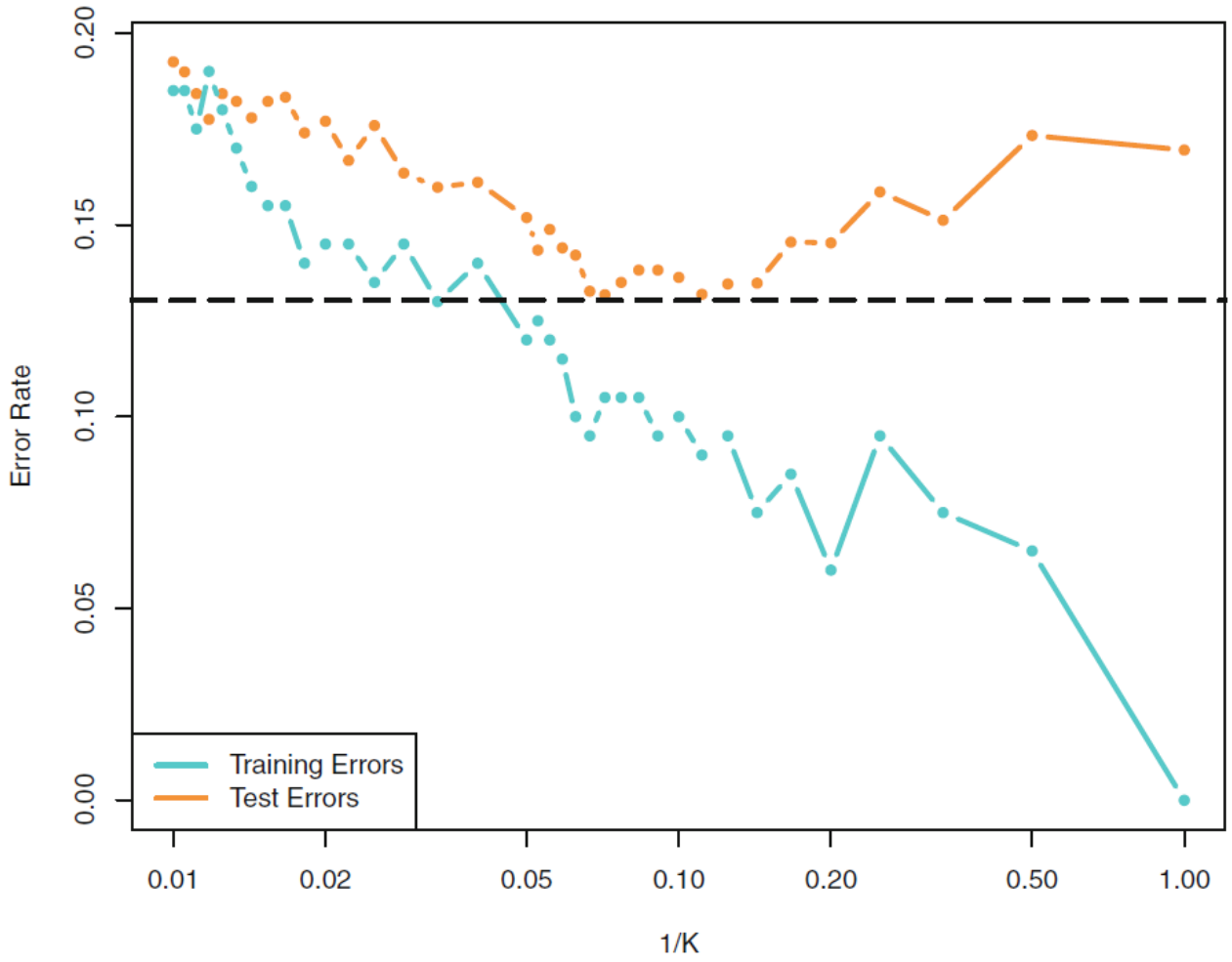




- 왼쪽 그림
  - An illustrative example of the KNN approach: make a prediction for the point labeled by the black cross.
  - Suppose that we choose $K = 3$.
  - KNN will first identify the three observations that are closest to the cross: two blue points and one orange point, resulting in estimated probabilities of 2/3 for the blue class and 1/3 for the orange class. Hence KNN will predict that the black cross belongs to the blue class.
- 오른쪽 그림
  - applied the KNN approach with $K = 3$ at all of the possible values for X_1 and X_2, and have drawn in the corresponding KNN decision boundary.
- KNN can often produce classifiers that are surprisingly close to the optimal Bayes classifier.
- The choice of K has a drastic effect on the KNN classifier obtained.
  - When $K = 1$, the decision boundary is overly flexible and finds patterns in the data that don't correspond to the Bayes decision boundary. $\rightarrow$ **low bias** but very **high variance**
  - As K grows, the method becomes less flexible and produces a decision boundary that is close to linear. $\rightarrow$ **low-variance** but **high bias**



- As $1/K$ increases, the method becomes more flexible. As in the regression setting, the training error rate consistently declines as the flexibility increases. However, the test error exhibits a characteristic U-shape.



- In both the regression and classification settings, choosing the correct level of flexibility is critical to the success of any statistical learning method.